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 $\geq$ 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:*

$$P(X \leq x; R, r) = F(x; R, r) = \sum_{k=1}^{x} f(k; R, r)$$
$$= 1 - \frac{\binom{R-x}{r}}{\binom{R}{r}}. \quad (5)$$

This can be used for assessing if the observed number of short distances is higher than would be expected by random chance. Proofs of the theorems are found in the Appendix section.

2.1. Hypothesis testing

As mentioned in the introduction, we are typically interested in if the successes are evenly distributed or if they are clustered somewhere in the series. It is natural to start out considering the population mean, which we denote as $\bar{X}$. For a given series of R events with r successes, $\bar{x}$ is the observed average distance between successes, which we may think of as a sample from $\bar{X}$. However, $\bar{x}$ contains very limited information about the clustering of successes. In fact, given R and r a sufficient statistic for $\bar{x}$ is the distance from the first to the last success.

A more informative and tunable test for clustering of successes involves the cumulative function from Eq. (5). Let the random variable Y be the number of distances smaller than x_{lim}, having realisation $y = \#(x < x_{lim})$. The expected value is $E(Y) = (r-1)F(x_{lim}; R, r) = \mu_{x_{lim}}$. We formulate the hypothesis test:

$$\begin{aligned} H_0 &: y = E(Y) \\ H_1 &: y > E(Y), \end{aligned} \quad (6)$$

where the alternative hypothesis is that the number of small distances are over-represented. As distances are either smaller than x_{lim} or not, Y can be described as binomial with $n = r-1$ trials, $k = y$ successes, and a probability of success equal to $p = F(x_{lim}; R, r)$. The associated p-value $P(Y \geq y) = \sum_{k=y}^{r-1} \binom{r-1}{k} p^{r-1}(1-p)^{r-1-k}$ is the probability of observing Y at least as large as y given that H_0 is true.

The critical value of the test is found using the quantile function for the binomial distribution with the same p and $n = r-1$ trials, and a quantile value of $1-\alpha$ (significance level). Correspondingly, the power of the test is found by calculating the cumulative binomial probability for the critical value with $n = r-1$ trials and a probability of success $p = y/(r-1)$.

3. Software description

The software package *fixedTimeEvents* is implemented in the programming language R. It follows the template used by the R language's distribution functions where the letters d, p, q and r at the beginning of a function name indicate the probability density (mass), cumulative distribution, quantile, and random number functions, respectively. For simplicity *Liland* is used as the distribution name in the R package, i.e. *dLiland*, *pLiland*, *qLiland* and *rLiland*. Installation of the package is done per usual R software practise either through user interface menus (if available) or by writing *install.packages('fixedTimeEvents')* in the console/terminal. Before describing the software in more details, a note has to be made on numerical limitations.

3.1. Limitations of the binomial coefficient

Binomial coefficients are impossible to compute directly when numbers become large, e.g. $\binom{4000}{2000}$ typically is reported as Inf by mathematical software. However, doing the calculations in the logarithmic domain and utilising for instance Stirling's approximation (~1730 AD) of the factorial (second order approximation in Eq. (7)), one can achieve a high degree of accuracy also for arbitrarily large problems. All calculations for the probability mass function are performed in the logarithmic domain before converting the result back through the exponential function.

$$\begin{aligned} \ln(n!) &\approx n\ln(n) + \ln\left(n\sqrt{\frac{2\pi}{n}}\right) - n + \ln\left(1 + \frac{1}{12n}\right) \\ \ln f(x; R, r) &\approx \ln((R-x)!) - \ln((r-1)!) \\ &\quad - \ln((R-x-r+1)!) \\ &\quad - \ln(R!) + \ln(r!) + \ln((R-r)!). \end{aligned} \quad (7)$$

In the software implementation we use the exact binomial coefficients when $\binom{R}{r} < 10^{20}$ to avoid large round-off errors and incalculable results and switch to the approximations otherwise to obtain the highest possible accuracy in all cases. An exception is made for the cumulative distribution function with approximation, where the maximum value $x = R - r + 1$ would lead to the calculation of $log(-1)$ because $\binom{R-(R-r+1)}{r} = \binom{r-1}{r}$. The result should be exactly 1 and is thus plugged directly into return value(s).

3.2. Implementation

In the R language, the binomial coefficient is represented by the function *choose*. As described in the previous subsection, we use the function *choose* whenever possible, i.e. when it returns a finite value $<10^{20}$, and replace this by Stirling's approximation otherwise.

The quantile function does not have an analytical solution. As the distribution is discrete, generating all possible distribution values for moderate sized R and searching using the *match* function finds the solution without noticeable lag. For $R > 10^6$ a sequential search is made, dividing the problem into manageable sequences of length 10^6 each and stopping as soon as the solution is found.

In addition to the main distribution functions, two more sets of functions are available in the package: statistical testing functions and simulation functions. The former are named *Liland.test*, *Liland.crit* and *Liland.pow*. These are the proposed test, the critical value of the test and the power of the test. The latter are named *simLiland*, *simLiland2* and *simLilandMu*. They perform sampling repeatedly from a given Liland distribution, sampling from the Bernoulli distribution and summarising, and sampling random mean "Liland" numbers.

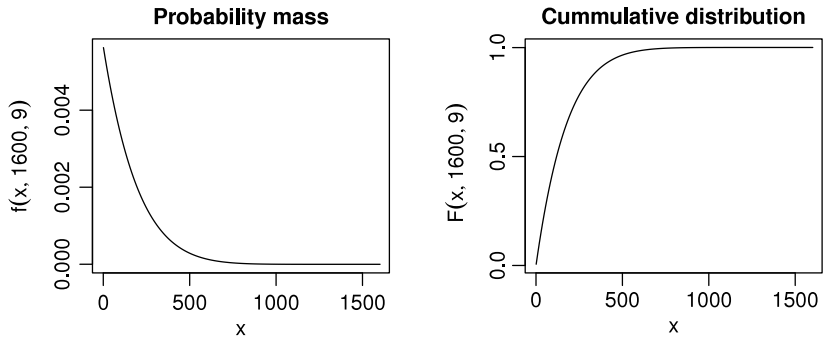

Fig. 1. Probability mass and cumulative distribution for $R = 1600$ and $r = 9$.

Table 1
All possible ways of distributing 3 successes among 5 events.

#	Position				
1	✓	✓	✓		
2	✓	✓		✓	
3	✓	✓			✓
4	✓		✓	✓	
5	✓		✓		✓
6	✓			✓	✓
7		✓	✓	✓	
8		✓	✓		✓
9		✓		✓	✓
10			✓	✓	✓

4. Illustrative examples

4.1. Counting occurrences

An example small enough to show all possible combinations of positions, while still shedding some light on the properties of the proposed distribution, is the one used previously with $R = 5$ and $r = 3$. In Table 1 all possible ways of distributing 3 successes among 5 trials are shown. We observe that there are 10 rows in the table corresponding to the $\binom{R}{r} = \binom{5}{3} = 10$ ways to sample 3 successes from 5. Also, there is a total of $(3-1)\cdot\binom{5}{3} = 20$ pairs of successes among these. There are three possible values for $x \in 1, 2, 3$. $x = 1$ means that a pair of consecutive successes are neighbours.

Starting with the largest value of x, we can count the occurrences of $x = 3$ in Table 1 to be once on row 3 and once on row 6. This verifies the expanded numerator $(r-1)\cdot\binom{R-3}{r-1} = 2\cdot\binom{2}{2} = 2$ for $x = 3$. For $x = 1$ and $x = 2$ the counts would be $2\cdot\binom{4}{2} = 12$ and $2\cdot\binom{3}{2} = 6$, respectively. Combining numerators and denominators we have: $P(X = 1; R = 5, r = 3) = \frac{12}{20} = \frac{6}{10}$, $P(X = 2; R = 5, r = 3) = \frac{6}{20} = \frac{3}{10}$, and $P(X = 3; R = 5, r = 3) = \frac{2}{20} = \frac{1}{10}$. These naturally sum to 1.

4.2. Sample distributions

To illustrate some properties of the proposed distribution we have plotted an example of the probability mass and cumulative distribution for $R = 1600$ and $r = 9$ in Fig. 1. These values of R and r could for instance be the number of bases of a 16S rRNA sequence and the number of hyper-variable regions in it [4]. We observe that the probability mass function is monotonically decreasing, and that it is most steep close to 1. The steepness will vary with the ratio R/r, but the general shape will be similar for all possible values of R and r.

```
R <- 1600    # trials
r <- 9       # successes

# Probability mass and distribution values for all
# possible values of x (vector from 1 to R-r+1):
mass         <- dLiland(1:(R-r+1), R, r)
distribution <- pLiland(1:(R-r+1), R, r)
```

A quick visual check of the distribution of ones data can be done by a quantile–quantile plot, like the one in Fig. 2. In this example 1000 observations have been sampled with $R = 2000$ and $r = 120$ and plotted against the theoretical distribution with the same properties. Here the observations mostly follow the qq-line (the line through quantiles 0.25 and 0.75), as expected. There is less coincidence towards the high values of x. This is also expected, since observations are less densely sampled here.

```
R <- 2000    # trials
r <- 120     # successes
n <- 1000    # random samples

# Random sampling and theoretical quantile
# values from the proposed distribution:
random.sample     <- sort(rLiland(n, R, r))
theoretical.distr <- qLiland(ppoints(n), R, r)
```

4.3. Gene regulation

We consider the case where 1949 genes are arranged along a bacterial chromosome in the same reading direction. Their position relative to each other and possible overlap is not interesting in this case, only the order of the start of the genes. In a study of gene expression, a total of 162 genes have been classified as regulated, based on some per-gene test procedure. We consider the status of each gene (regulated/not regulated) as a Bernoulli trial, where 'regulated' is the same as 'success', i.e. $R = 1949$ and $r = 162$. Bacterial genes are often arranged in operons, i.e. several neighbouring genes are, more or less, regulated by the same sigma factor (see e.g. [1]). This means

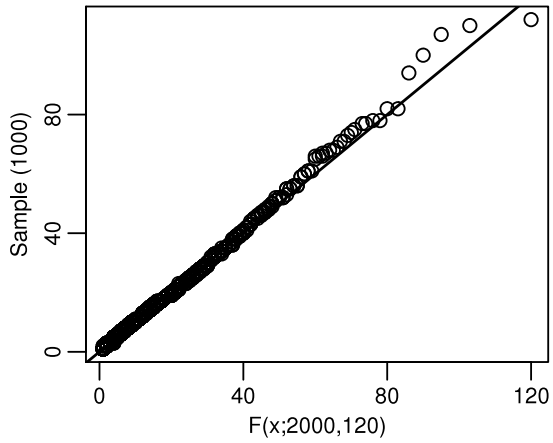

Fig. 2. Quantile–quantile plot of 1000 randomly sampled observations with $R = 2000$ and $r = 120$ against the theoretical distribution (qq-line in dashed line type).

that regulated genes should show a tendency to cluster along the chromosome, unless the expression data and subsequent testing procedure have allowed too many genes to obtain a false positive status as regulated. In the latter case we expect the (apparently) regulated genes to appear evenly distributed.

This corresponds to a hypothesis test situation as in Eq. (6), where we want to test if the distance $x = 1$ occurs more frequently than we would expect if the regulated genes were distributed at random, i.e. $x_{lim} = 2$. From our proposed cumulative distribution function in Eq. (5) we get $F(1; 1949, 162) = 0.083$ and $E(Y) = (162 - 1)F(1; 1949, 162) = 13.4$. Using 0.083 as the binomial probability, we easily find that if at least 19 of the $r - 1 = 161$ distances have the value 1, we can reject H_0 of Eq. (6) at a 5% level, indicating that genes classified as regulated tend to be neighbours. Looking at Fig. 3, we see that the observed $y = 73$ is far to the right of the critical values and thus associated with a very low p-value. This gives support to the claim that the regulated genes tend to be neighbours.

```
R <- 1949   # trials
r <- 162    # successes

# P(x < 2; R = 1949, r=162):
p <- pLiland(1, R, r)           # = 0.083

# Probability of randomly observing
# y > 1, ..., 80, the corresponding
# critical value and power at y=73:
Prob <- pbinom(1:80, r-1, p,
               lower.tail = FALSE)
crit <- Liland.crit(1, R, r)    # = 19
pow  <- Liland.pow(1, R, r, 73) # ≈ 1
```

4.4. Lottery numbers

Lottery numbers are sometimes a theme of conversation, especially with regard to the apparent over-representation of consecutive numbers that are often drawn. In a lottery where r numbers are drawn from R possible outcomes, e.g. $r = 7$ and $R = 34$, the distance between the drawn numbers (after sorting) follows the proposed distribution. If we consider a weekly

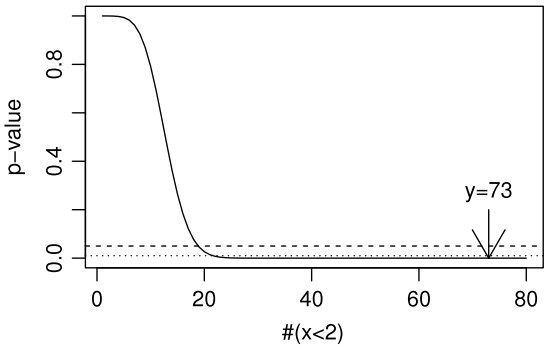

Fig. 3. Probability of observing more distances shorter than 2 ($P(Y > y|x < 2)$) than the expected number $E(Y) = 13$ as a function of the observed number (y). Critical values for 5% and 1% test level are 19 and 22, respectively.

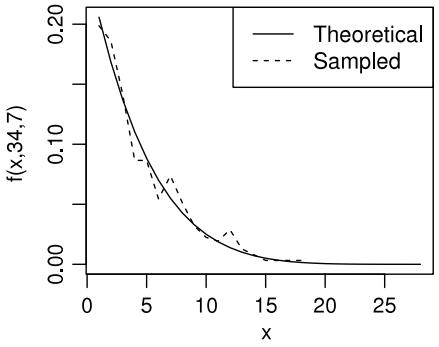

Fig. 4. Probability mass for $R = 34$ and $r = 7$, corresponding to drawing 7 lottery numbers from 34 possible outcomes. The dashed line indicates the corresponding proportions from 52 weekly random draws.

lottery we can sample uniformly 7 numbers from $[1, 2, \ldots, 34]$ without replacement 52 times to obtain the 52 x 6 distances that would constitute a normal year of lottery. In Fig. 4 the proportions of each distance from such an experiment is plotted together with the theoretical probability mass.

```
# Sample 52 sets of lottery numbers and
# put them in the matrix X:
X <- matrix(0, 52, 7)
for(i in 1:52)
  X[i, ] <- sort(sample(1:34, 7, replace = FALSE))

# Calculate the distance between successes
# in each of the lottery draws:
Xdiff <- t(apply(X, 1, diff))

# Calculate the proportions of occurrences
# for each possible distance between successes
# and calculate the theoretical mass values
# from the proposed distribution:
proportions <- table(c(Xdiff)) / (52*6)
theoretical.distr <- dLiland(1:28, 34, 7)
```

This confirms the typical notion that lottery numbers are often consecutive, as 20.6% of the distances are 1, and more than 50% have distances of $x \leq 3$. The expected number of pairs of consecutive numbers in a single draw is $(7 - 1) \cdot F(1; 34, 7) = 6 \cdot 0.206 = 1.235$. Critical values for the proposed test for over-

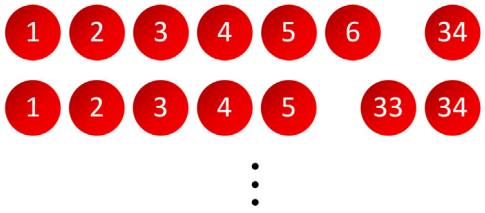

Fig. 5. Lottery numbers clustering around the extreme values 1 and 34. These are 2 of the 6 possible draws that would lead to a distance between consecutive numbers of 28.

representation of neighbouring pairs at 5% and 1% levels of significance are at 3 and 4, respectively. This means that 3 or 4 (possibly consecutive) pairs of consecutive numbers would happen regularly, just by chance, while 5 pairs would be more unlikely (critical value for 0.1% level of significance). The latter would occur only if all numbers were clustered in two consecutive groups. The least probable type of draw we can observe is $P(X = 28; R = 34, r = 7) = 6/\left(6 \times \binom{34}{7}\right) = 1/\binom{34}{7} = 1.86 \times 10^{-7}$, corresponding to series of lottery numbers clustered near the ends as shown in Fig. 5.

5. Conclusions

In this article we have introduced theory and an R package for the exact distribution of distances between discrete events in fixed time. As shown, the distribution is easily verifiable for small, countable examples. For larger samples it can be approximated by a discretised version of the exponential distribution, but using Stirling's approximation of the factorial instead, we can achieve high precision for problems of all sizes.

Testing concerning the mean distance turns out to be of little value as this would only test for the combined distance from the first to the last success, which seldom is of interest. However, testing using the binomial distribution can be applied, e.g. to see if short distances are over- or under-represented. The gene regulation example shows how the proposed distribution can be applied on real data, while the lottery example is an elegant demonstration of why lottery numbers often cluster.

The R package *fixedTimeEvents* includes all basic distribution functions, statistical testing functions and functions to perform simple, numerical simulations regarding the distribution. The main motivation for generating the distribution and accompanying test of over-representation of short distances has been problems in bioinformatics, but the theory is fully general for any problem of equal characteristics. Implementing this in the R language, makes it available directly for R users and indirectly for users of languages that can call upon R libraries like Python, e.g. [5].

Appendix. Proofs

Proof of Theorem 1. The mean value of a discrete distribution is $\mu = E(X) = \sum_x x \times P(X = x)$. If we write this out using the proposed distribution and apply the relationship $\sum_{j=k}^{n}(n+1-j)\binom{j-1}{k-1} = \binom{n+1}{k+1}$ we get:

$$E(X) = 1\frac{\binom{R-1}{r-1}}{\binom{R}{r}} + 2\frac{\binom{R-2}{r-1}}{\binom{R}{r}} + \cdots + (R-r+1)\frac{\binom{r-1}{r-1}}{\binom{R}{r}}$$

$$= \frac{\sum_{j=1}^{R-r+1} j\binom{R-j}{r-1}}{\binom{R}{r}} = \frac{\binom{R+1}{r+1}}{\binom{R}{r}} = \frac{R+1}{r+1}. \tag{A.1}$$

Proof of Theorem 2. The cumulative distribution is the sum of all masses up to $X = x$, i.e.:

$$F(X = x; R, r) = \sum_{k=1}^{x} \frac{\binom{R-k}{r-1}}{\binom{R}{r}}$$

$$= \frac{\binom{R-1}{r-1} + \binom{R-2}{r-1} + \cdots + \binom{R-x}{r-1}}{\binom{R}{r}}. \tag{A.2}$$

If we add and subtract $\binom{R-x}{r}$ from the numerator, we can use the recursive formula $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$ to sequentially collapse the sum:

$$\frac{\binom{R-1}{r-1} + \cdots + \binom{R-x}{r-1} + \binom{R-x}{r} - \binom{R-x}{r}}{\binom{R}{r}}$$

$$= \frac{\binom{R-1}{r-1} + \cdots + \binom{R-x+1}{r-1} + \binom{R-x+1}{r} - \binom{R-x}{r}}{\binom{R}{r}}$$

$$\vdots$$

$$= \frac{\binom{R}{r} - \binom{R-x}{r}}{\binom{R}{r}} = 1 - \frac{\binom{R-x}{r}}{\binom{R}{r}}. \tag{A.3}$$

Proof of Theorem 3. The recurrence relationship between $P(X = x; R, r)$ and $P(X = x+1; R, r)$ can be shown by a simple factorisation of the numerator since the denominator is unchanged:

$$\binom{R-x}{r-1} = \frac{(R-x)!}{(r-1)!(R-x-(r-1))!}$$

$$= \frac{(R-(x+1))!(R-x)}{(r-1)!(R-(x+1)-(r-1))!(R-x-(r-1))}$$

$$= \binom{R-(x+1)}{r-1}\frac{R-x}{R-x-(r-1)}. \tag{A.4}$$

Solving for $\binom{R-(x+1)}{r-1}$ and reintroducing the denominator we get the following:

$$\frac{\binom{R-(x+1)}{r-1}}{\binom{R}{r}} = \frac{R-x-(r-1)}{R-x}\frac{\binom{R-x}{r-1}}{\binom{R}{r}}$$
$$= \left(1-\frac{r-1}{R-x}\right)\frac{\binom{R-x}{r-1}}{\binom{R}{r}}. \quad \text{(A.5)}$$

References

[1] Diep DB, Straume D, Kjos M, Torres C, Nes IF. Peptides 2009;30:1562–74.

[2] Forbes C, Evans M, Hastings N, Peacock B. Statistical distributions. fourth ed. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2000.

[3] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2016. Available from: http://www.R-project.org/.

[4] Van de Peer Y, Chapelle S, Wachter RD. Nucleic Acids Res 1996;24:3381–91.

[5] Xia XQ, McClelland M, Wang Y. J Stat Softw 2010;35.