Available online at www.sciencedirect.com
ScienceDirect
BioSystems 90 (2007) 509–515
www.elsevier.com/locate/biosystems

Parameter-free testing of the shape of a probability distribution

M. Broom[a], P. Nouvellet[b], J.P. Bacon[b], D. Waxman[b,*]

[a] *School of Science and Technology, University of Sussex, Falmer, Brighton BN1 9QG, Sussex, UK*
[b] *School of Life Sciences, University of Sussex, Falmer, Brighton BN1 9QG, Sussex, UK*

Received 30 August 2006; received in revised form 5 December 2006; accepted 6 December 2006

Abstract

The Kolmogorov–Smirnov test determines the consistency of empirical data with a particular probability distribution. Often, parameters in the distribution are unknown, and have to be estimated from the data. In this case, the Kolmogorov–Smirnov test depends on the form of the particular probability distribution under consideration, even when the estimated parameter-values are used within the distribution. In the present work, we address a less specific problem: to determine the consistency of data with a given functional form of a probability distribution (for example the normal distribution), without enquiring into values of unknown parameters in the distribution. For a wide class of distributions, we present a direct method for determining whether empirical data are consistent with a given functional form of the probability distribution. This utilizes a transformation of the data. If the data are from the class of distributions considered here, the transformation leads to an empirical distribution with no unknown parameters, and hence is susceptible to a standard Kolmogorov–Smirnov test. We give some general analytical results for some of the distributions from the class of distributions considered here. The significance level and power of the tests introduced in this work are estimated from simulations. Some biological applications of the method are given.

Keywords: Kolmogorov–Smirnov test; Non-parametric method; Functional form of a distribution

1. Introduction

A standard method of testing whether measured data are consistent with a particular continuous probability distribution is the Kolmogorov–Smirnov (KS) test (see standard statistical texts, for example Hogg and Tanis, 2006). The essence of the test is to compare the maximum distance, termed D_{KS}, between the empirical cumulative distribution and the particular cumulative distribution of interest. In many practical cases, the distribution of interest contains one or more parameters that are not known *a priori*, but have to be estimated from the data. When this applies, the distance statistic, D_{KS}, that underlies the KS test, no longer has a universal distribution, i.e. one that is independent of the distribution under consideration. The procedure introduced by Lilliefors (1969) to test an empirical distribution, when parameters have to be determined from the data, is to carry out numerical simulations and thereby produce the distribution of the test statistic that is relevant for the particular distribution under consideration.

Here we consider a less specific question than directly testing whether measured data are consistent with a particular distribution. Our aim is to determine whether the measured data are consistent with the general *functional form* (i.e. shape) of a cumulative distribution of interest, independent of the value of any unknown parameters upon which the distribution depends. Thus, we might wish to test whether our data could consistently be interpreted as arising from an exponential distribution, without having any interest in the value of the single parameter characterising the exponential dis-

[*] Corresponding author. Tel.: +44 1 273 678 559; fax: +44 1 273 678 937.
E-mail address: D.Waxman@sussex.ac.uk (D. Waxman).

0303-2647/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2006.12.002

Table 1
Lists some of the distributions arising from random variables X of the form $X = a + b\xi$, where a is the location parameter, b the scale parameter and ξ is the standard random variable

Distribution	Location parameter	Scale parameter	Standard random variable ξ
1	α	$\beta - \alpha$	U
2	0	λ	$-\log(U)$
3	0	λ	$\pm\log(U)$
4	μ	σ	Z
5	α	β	e^Z

For a wide range of important probability distributions (including those mentioned in this paper), the functional forms are given in texts on probability theory or mathematical statistics (such as Hogg and Tanis, 2006 or Weiss, 2006). Some of the distributions in this table have standard names and symbols: (1) is the uniform distribution and written as $U[\alpha, \beta]$; (2) is the exponential distribution and written as $\exp(\lambda)$; (3) with the sign, ± picked at random with equal probability, we have a reflected exponential distribution; (4) is a normal distribution and written as $N(\mu, \sigma^2)$; (5) corresponds to a particular case of a lognormal distribution that has been horizontally translated by an amount α. We shall simply refer to it as a lognormal distribution in the present work.

tribution. Focussing solely on the functional form of the distribution means our test is non-parametric in nature.

The approach adopted here allows a direct test of the data when the random variables underlying the distribution of interest may be expressed in the form:

$$X = a + b\xi \tag{1}$$

where a is the *location* parameter, b the non-zero *scale* parameter and ξ is the *standard* random variable (i.e. a random variable from a known distribution, with no unknown parameters). Before we proceed with the derivation of our results, we note that there are a number of well-known examples of random variables that may be defined by an expression of the form of Eq. (1). In terms of $U \equiv U[0, 1]$, a random number that is uniformly distributed on [0, 1], and Z, a standard normal random variable (i.e. with mean zero and variance unity), some examples of random variables of the form in Eq. (1) are given in Table 1.

2. Methodology

We shall consider the general problem shortly, but first investigate the special case where the location parameter of the distribution, a, is zero.

2.1. Method 1: location parameter is zero

The simplest case arises when the location parameter a has the value $a = 0$. This is the case where the particular distribution under consideration is specified by a single parameter, b. The shape of the distribution of the random variable X, of Eq. (1), is determined from the distribution of the standard random variable ξ. Denoting the probability density of ξ by $f_\xi(x)$ and its cumulative distribution by $F_\xi(x)$, the probability density and cumulative distribution of $X \equiv b\xi$ are $f_\xi(x/b)/b$ and $F_\xi(x/b)$, respectively. These evidently depend upon the parameter b. We can derive a distribution related to that of X that is independent of the parameter b as follows.

Let X_i and X_j represent independent draws $(i \neq j)$ of $X = b\xi$ from its distribution and let us use the notation $|X|_< = \min(|X_i|, |X_j|)$ and $|X|_> = \max(|X_i|, |X_j|)$. We show in Appendix A that the cumulative distribution and probability density of:

$$R_0 = \frac{|X|_<}{|X|_>} \tag{2}$$

are, for r in the range $1 \geq r \geq 0$, given by:

$$F_{R_0}(r) = 2\int_{-\infty}^{\infty} f_\xi(y)[F_\xi(r|y|) - F_\xi(-r|y|)]\,\mathrm{d}y \tag{3}$$

$$f_{R_0}(r) = 2\int_{-\infty}^{\infty} |y| f_\xi(y)[f_\xi(r|y|) + f_\xi(-r|y|)]\,\mathrm{d}y \tag{4}$$

Outside the range $1 \geq r \geq 0$, $f_{R_0}(r)$ vanishes, hence $F_{R_0}(r)$ vanishes for $r < 0$ and is unity for $r > 1$. Because the parameter b cancels between the numerator and denominator in Eq. (2), it follows that R_0 is independent of any parameters and hence $F_{R_0}(r)$ is determined solely by the parameter-free cumulative distribution of the standard random variable ξ, namely $F_\xi(x)$. A standard KS test may then be directly applied, since $F_{R_0}(r)$ is a known distribution with no unknown parameters. Thus, we can proceed by testing whether the empirical distribution of the data is consistent with $F_{R_0}(r)$ *without* having to first determine any parameters from the data.

In practical applications, there remains the question of how to construct the set of values of the R_0 statistic from the set of realised X values, i.e. from the set $\{x_1, x_2, \ldots, x_N\}$, where N is the number of measured values. Assuming N is even (or is made so, by discarding the final x), we use each x value only once, by forming $N/2$ ratios, for example $\{x_1/x_2, x_3/x_4, \ldots, x_{N-1}/x_N\}$. These are converted to R_0 values by (i) taking their absolute value and then (ii) taking the smaller of the resulting absolute value or its reciprocal. For example, the corresponding R_0 value obtained from x_1 and x_2 is the smaller of $|x_1/x_2|$ and $|x_2/x_1|$.

In the discussion, we give the rationale for the particular form of the R_0 statistic adopted.

2.2. Method 2: general case

It is straightforward to observe that in the general case, where $a \neq 0$, the quantity $(X_i - X_j)/(X_k - X_l)$ equals $(\xi_i - \xi_j)/(\xi_k - \xi_l)$ and hence is independent of a and b. Accordingly we define:

$$R = \frac{|X - X|_<}{|X - X|_>} \tag{5}$$

where $|X - X|_< = \min(|X_i - X_j|, |X_k - X_l|)$ and $|X - X|_> = \max(|X_i - X_j|, |X_k - X_l|)$ for i, j, k and l all different.

Following the same logic as in Method 1 (see Appendix A), the cumulative distribution and probability density of R are, for $1 \geq r \geq 0$, given by:

$$F_R(r) = 2\int_{-\infty}^{\infty} f_{\xi-\xi}(y)[F_{\xi-\xi}(r|y|) - F_{\xi-\xi}(-r|y|)]\,\mathrm{d}y \tag{6}$$

$$f_R(r) = 2\int_{-\infty}^{\infty} |y| f_{\xi-\xi}(y)[f_{\xi-\xi}(r|y|) + f_{\xi-\xi}(-r|y|)]\,\mathrm{d}y \tag{7}$$

where $f_{\xi-\xi}(\bullet)$ is the probability density of $\xi_i - \xi_j$, i.e. $f_{\xi-\xi}(x) = \int_{-\infty}^{\infty} f_\xi(y) f_\xi(y - x)\,\mathrm{d}y$, and $F_{\xi-\xi}(\bullet)$ is the corresponding cumulative distribution. Outside the range $1 \geq r \geq 0$, $f_R(r)$ vanishes, hence $F_R(r)$ vanishes for $r < 0$ and is unity for $r > 1$.

By construction, the cumulative distribution of R, i.e. $F_R(r)$, and its probability density, $f_R(r)$, are independent of unknown parameters present in the probability density of X, and hence as in Method 1, a standard KS test may be directly employed.

In this more general case, we again use each X data value only once in the construction of the set of R values and the simplest way to proceed, assuming N is divisible by 4 (or is made so by discarding the necessary final few x's) is to form a set of $N/4$ values of R given by $\{(x_1 - x_2)/(x_3 - x_4), (x_5 - x_6)/(x_7 - x_8), \ldots, (x_{N-3} - x_{N-2})/(x_{N-1} - x_N)\}$. We then, again, take the absolute value of these and select the smaller of the resulting absolute value or its reciprocal.

3. Application to some standard distributions

Details of the calculations are given in Appendix A.

We note that by definition of R_0 or R, these random variables lie in the range [0, 1] and, as a result, all probability densities are non-zero only for r in the range [0, 1]. The corresponding cumulative distributions are zero for $r < 0$ and unity for $r > 1$. The results given in this section will only apply for the range of r where the cumulative distributions exhibit non-trivial behaviour, namely $1 \geq r \geq 0$, without this restriction being further stated in the results.

3.1. Exponential distribution

First consider the case where $a = 0$ and $\xi \sim \exp(1)$, so $f_\xi(x) = \exp(-x)$ for $x \geq 0$ and vanishes otherwise.

Method 1 (above) leads, via Eq. (3), to the cumulative distribution of R_0 being given by $F_{R_0}(r) = 2r/(1 + r)$ and to the probability density $f_{R_0}(r) = 2/(1 + r)^2$.

Method 2 (above), for $a \neq 0$, yields, via Eq. (6), to identical results to those of Method 1, i.e. to $F_R(r) = 2r/(1 + r)$ and $f_R(r) = 2/(1 + r)^2$.

3.2. Reflected exponential distribution

For the reflected exponential distribution, $f_\xi(x) = \exp(-|x|)/2$, Method 1 leads, via Eq. (3), to the cumulative distribution of R_0 being given by $F_{R_0}(r) = 2r/(1 + r)$ and to the probability density $f_{R_0}(r) = 2/(1 + r)^2$.

Method 2 leads, via Eq. (6), to a cumulative distribution of R being given by $F_R(r) = (r/2)(3 + 9r + 4r^2)/(1 + r)^3$ and to the probability density $f_R(r) = (3/2)(1 + 4r + r^2)/(1 + r)^4$.

3.3. Normal distribution

We have $\xi \sim N(0, 1)$ so $f_\xi(x) = \exp(-x^2/2)/\sqrt{2\pi}$. Method 1 leads, via Eq. (3), to the cumulative distribution of R_0 being given by $F_{R_0}(r) = (4/\pi)\arctan(r)$ and to the probability density $f_{R_0}(r) = 4/[\pi(1 + r^2)]$. Method 2, which applies when $a \neq 0$, leads to identical results to those of Method 1: $F_R(r) = (4/\pi)\arctan(r)$ and $f_R(r) = 4/[\pi(1 + r^2)]$.

3.4. Uniform distribution

Consider first the case with $a = 0$. Using Method 1, the cumulative distribution of R_0 is given by $F_{R_0}(r) = r$ (i.e. coinciding with that of a uniform distribution on [0, 1]) and to the probability density $f_{R_0}(r) = 1$. For the case where $a \neq 0$ we use Method 2, and obtain a cumulative distribution of R given by $F_R(r) = (r/3)(4 - r)$ and to the probability density $f_R(r) = (2/3)(2 - r)$.

4. Numerical tests

In this section, we give results from simulated data from various distributions, and compare the simulated data to the distributions $F_{R_0}(\bullet)$ or $F_R(\bullet)$ evaluated above. We performed two sets of comparisons, first using Method 1 and then Method 2.

4.1. Method

To evaluate the characteristics of the statistical tests proposed here, we investigated the associated errors. We note that it is usual to distinguish two kinds of error:

(i) A type I error corresponds to rejecting a valid null hypothesis. The probability of a type I error is termed the *significance level* of the test.
(ii) A type II error corresponds to accepting an invalid null hypothesis. The complement of the probability of a type II error gives the *power* of the test.

A good statistical test should have low probabilities of producing both type I and type II errors.

Generally, to perform a KS test, one estimates the maximum distance of the cumulative distribution derived from observation, $F_{\text{obs}}(\bullet)$, from the theoretical one, such as $F_R(\bullet)$. This distance, denoted D_{KS}, is known as the Kolmogorov–Smirnov statistic, and is defined as $D_{\text{KS}} = \max_x |F_{\text{obs}}(x) - F_R(x)|$. Using D_{KS}, it is possible to assess if an observed distribution is effectively indistinguishable from a theoretical distribution (this is the null hypothesis). Thus, if the distance between the distributions, D_{KS}, is greater than the critical distance associated with the KS test, for the sample size under consideration, then the null hypothesis is rejected. For our purposes we use a p value of 0.05. We thus reject the null hypothesis when a value of D_{KS} is obtained that would arise, by chance, in $\leq 5\%$ of the cases where the null hypothesis is valid.

One of the major constraints of the KS test, as explained above, is that any parameters appearing in the theoretical distribution must be known. The change in variable adopted here (from X to R_0 or R; see Eqs. (2) and (5)), allows us, in a number of cases of interest, to test if a random variable follows a theoretical distribution of given functional form (i.e. shape), independent of any parameters in the distribution of X.

To find the significance level of the test (the probability of type I errors), we generated random numbers from exponential, normal and uniform distributions. These were used in Methods 1 and 2 (for the variables R_0 and R described above). We evaluated the distance, D_{KS}, between the cumulative distribution of a simulated sample and the particular distribution used to generate the sample. The null hypothesis was listed as rejected at the 5% level when D_{KS} was greater than the critical value of this statistic for the sample size adopted. Critical values of D_{KS} follow from a standard KS test and can be found in published tables. Any such rejections correspond to the false conclusion that the sample comes from a distribution different from the specified theoretical distribution, and indicate a type I error of the test.

We next generated random numbers from lognormal and reflected exponential distributions. These distributions are, respectively, very similar in shape to exponential and normal distributions. The data from these simulations were tested to see if:

(i) The sample drawn from a lognormal distribution is effectively indistinguishable from a theoretical exponential distribution.
(ii) The sample drawn from a reflected exponential distribution is effectively indistinguishable from a theoretical normal distribution.

By evaluating D_{KS} we estimated the probability of (i) or (ii) being *accepted* as an indication of type II errors i.e. as an indication of the power of the test.

For each comparison, we computed D_{KS} for different sample sizes, N, ranging from $N = 10$ to 2000. For each sample size, v replicate samples were generated and v was chosen so that $v \times N = 3 \times 10^6$ was the total number of random numbers generated. The D_{KS} values obtained were used to estimate the probability of acceptance of the null hypothesis (type II errors), for each comparison of distributions, and for each sample size. Our estimates of the various errors are summarised in Table 2.

5. Results

5.1. The significance level of the test

We obtained an estimate of the significance level of the test from our data by taking a large number of samples from a specified distribution and finding the proportion where the null hypothesis was erroneously rejected. As we were using the standard KS test with a significance level of 5% we would naturally expect an estimate in the region of 5% in each case. We did this for the exponential, normal and uniform distributions and as expected, we did indeed, find that approximately 5% of the sample distributions led to rejection of the null hypothesis.

5.2. Power of the test

We investigated the power of the test by looking at the proportion of cases where the null hypothesis was incorrectly accepted. This was carried out for sample

Table 2
A summary of type I and type II errors is given for various distributions

Sample size	Estimate of the probability of type I error						Estimate of the probability of type II error			
	Uniform (e) vs. uniform (t)		Exponential (e) vs. exponential (t)		Normal (e) vs. normal (t)		Lognormal (e) vs. exponential (t)		Reflected exponential (e) vs. normal (t)	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
8	0.050	0.050	0.050	0.050	0.050	0.050	0.952	0.943	0.929	0.947
20	0.051	0.049	0.050	0.049	0.050	0.050	0.937	0.942	0.905	0.944
48	0.051	0.049	0.050	0.053	0.049	0.050	0.902	0.933	0.842	0.937
100	0.050	0.051	0.050	0.051	0.050	0.050	0.824	0.920	0.730	0.927
200	0.047	0.048	0.051	0.052	0.052	0.051	0.650	0.892	0.523	0.900
500	0.049	0.046	0.047	0.047	0.050	0.044	0.207	0.804	0.150	0.832
752	0.051	0.050	0.050	0.052	0.045	0.050	0.044	0.740	0.041	0.769
1000	0.057	0.042	0.050	0.047	0.051	0.049	0.008	0.666	0.013	0.713
2000	0.049	0.046	0.050	0.057	0.051	0.044	0.000	0.396	0.000	0.467

The null hypothesis is that "the observed distribution is drawn from the theoretical distribution". The label e or t following the name of a distribution in this table serves to distinguish between a distribution that is of effectively empirical (e) or theoretical (t) origin. Thus, X values are drawn from the distributions labelled e, and are our simulation of experimental data. The theoretical distribution of the variables R_0 or R (Eqs. (2) and (5)) are determined from the distribution labelled t and some examples of these distributions are given in Section 3. Under Method 1, all of the distributions labelled e, that were used to generate simulated X data, had a location parameter of $a = 0$ and a scale parameter of $b = 3$. The simulated X data were then transformed to R_0 data, according to Eq. (2) and a standard KS test was then employed (as described in the main text). Under Method 2, all distributions labelled e had a location parameter of $a = 2$ and a scale parameter of $b = 3$. The simulated X data were then transformed to R data, according to Eq. (5) and a standard KS test was, again, then employed. We used the tables of Neave (1989) for the critical values of the Kolmogorov–Smirnov statistic.

distributions that were similar in shape to the theoretical distribution that was used for the null hypothesis. We thus tested (i) a lognormal sample distribution (distribution 5 in Table 1) against a theoretical exponential distribution (distribution 2 in Table 1) and (ii) a reflected exponential sample distribution (distribution 3 in Table 1) against a theoretical normal distribution (distribution 4 in Table 1). We found that the test of Method 1 is, for a given sample size, sometimes less powerful than that of a standard KS test and sometimes more powerful. Method 2 performed less well than the test of Method 1. Since differences of two random variables are typically closer to being normally distributed than a single random variable, this may provide partial explanation of the relative performances of Methods 1 and 2, rather than attributing it solely to the reduction of the size of the data set of Method 1 (size $N/2$), to that of Method 2 (size $N/4$).

6. Biological applications of the method

The method presented in this paper can have a wide range of applications in biology and other scientific fields. Here we give as examples some possible behavioural ecology applications.

The method of parameter-free testing of the shape of a probability distribution was motivated from an analysis of the foraging behaviour of ants. In an experiment (Nouvellet et al., in preparation), times were recorded at which individual ants left their nest to explore a new area. The ants that left the nest, and arrived in the new area, were not able to return or communicate in any way with their nest mates. The distribution of *time intervals* that developed over the course of the experiment between each successive ant leaving the nest was calculated. There is the strong possibility that the distribution of time intervals will depend on the actual duration of the experiment. This follows since ants cannot, during the course of the experiment, return to the nest; hence the number within the nest declines over time and the rate at which ants leave the nest may also decline. Thus any distribution, for example, that characterising time intervals, may change over time; in other words any parameters it depends upon may be time dependent. We can however test if the *functional form* (or shape) of the distribution is time independent, if we make the assumption that during a typical time interval between two ants leaving the nest, any time-dependent parameters within the distribution change negligibly. Given this, we used Method 1 of this work to test whether the time intervals between ants leaving the nest had the *functional form* of an exponential distribution, with all time dependence carried by the single parameter characterising the exponential distribution. Thus, Method 1 (as fully described above) allows the use of the Kolmogorov–Smirnov test on the data to test whether the time intervals are exponentially distributed, without knowledge of the parameter in the distribution. Full details of the experiment and related experiments along with a detailed analysis of the data will be presented elsewhere (Nouvellet et al., in preparation).

This specific analysis can be extended, for example, to any type of social behaviour involving the movement of numbers of individuals. Examples are bees, wasps or termites leaving the nest or arriving at a food or water source. Additionally the variation of parameter(s) can be subsequently inferred to understand the manner in which it varies. Hence we can decouple different phenomena—the purely statistical from the values of parameters and their variation, using the methods of this paper.

It is clear that any behaviour that approximately repeats over time, and which might be influenced by external factors, could be analysed in this way to test whether there is a distribution of constant shape underlying the phenomenon. External factors could be the size of the 'source population' in the nest, the quantity or quality of the 'sink food source', but also environmental factors such as day-length.

On this last topic, we note that many temporal behaviours (hunting, singing, diving, onset of activity) appear to show a constantly shaped distribution that is shifted by the timing of an external event. Birds sing at sunrise (Fisler, 1962), but each day sunrise occurs at a different time, thus preventing the direct pooling of data taken at different times of year. Similarly, hunting behaviour (Van Orsdol, 1984), basking behaviour (Ciofolo and Boissier, 1992), onset of activity (Aschoff, 1966; Semenov et al., 2000) often occurs at certain times that exhibit variation over the course of a year.

There is thus a wide range of behavioural studies that could utilize the statistical approach described in this paper, where the functional form of a distribution is important and needs to be established.

7. Discussion

In this paper, we have introduced a method of data transformation which allows us to test whether data come from a particular functional form of a distribution, irrespective of the values of unknown parameters in the distribution. The distribution is of the form of Eq. (1) with ξ following a fully specified distribution. Many distributions can be written in the form of Eq. (1), and such a test is consequently of interest for a wide variety of practical situations.

For the transformed data, as given in Eqs. (2) and (5), we have specified the precise form of the distribution for several important standard distributions. Thus for these distributions, the method may be directly applied without further calculation. We have also given general expressions for the form of the probability density and cumulative distribution of the transformed data, in terms of the density and distribution of the underlying random variable, ξ, so that such expressions can be derived for other cases.

We have examined the power of our test in a variety of situations to examine the usefulness of the method. There does not appear to be a consistent advantage or disadvantage of Method 1, compared with the original Kolmogorov–Smirnov test for comparable data, where the distribution is known. For example, if in a standard KS test, we wish to discriminate between (i) data which are lognormally distributed, i.e. distribution 5 of Table 1 with $\alpha = 0$ and $\beta = 1$, with (ii) an exponential distribution (distribution 2 of Table 1, with $\lambda = 1$) we then find that it requires a sample size of $N \simeq 100$ so the different distributions can be discriminated in approximately 95% of all cases. By contrast, Method 1 of the present work requires $N \simeq 750$ to discriminate the functional form of the two distributions to the same level of power. However, applying the same procedure to (i) the reflected exponential distribution (distribution 3 of Table 1, with $\lambda = 1$) and (ii) a normal distribution (distribution 4 of Table 1, with $\mu = 0$ and $\sigma = 1$), a standard KS test requires a sample size of $N \simeq 850$ to discriminate the distributions in approximately 95% of all cases. By contrast, Method 1 of the present work requires $N \simeq 700$ to discriminate the functional form of the two distributions to the same level of power.

The power from the more general Method 2 is noticeably less than Method 1 (see Table 2) and thus there is a corresponding need for large data sets, which means that its range of applicability will be restricted.

In Section 2.1 of this work, we adopted a particular function of X_i/X_j, namely that of Eq. (2), as the statistic at the heart of Method 1, and a related quantity for Method 2. The virtue of the choice made is that R_0 lies in a compact range, whereas e.g. X_i/X_j can cover a very wide range of values, caused by potentially small denominators. We have found significantly different powers of a test based on R_0 or X_i/X_j; the wider range of the test statistic being accompanied by a significantly lower power to discriminate between different distributions.

The original motivation of this work arose from analysing the foraging behaviour of ants, as described in the previous section. We envisage that there may be a substantial range of other applications of the statistical methods presented here.

Acknowledgement

It is a pleasure to thank John Welch for helpful discussions.

Appendix A

In this appendix, we provide some details of the calculations underlying results in the main text.

First we find the general form for the cumulative distribution in Method 1. We can write $F_{R_0}(r) = \text{Prob}(|\xi_1|/|\xi_2| < r \,|\, |\xi_1| < |\xi_2|) = \text{Prob}(|\xi_1| < r|\xi_2| \,|\, |\xi_1| < |\xi_2|)$. For $r > 1$ we have $F_{R_0}(r) = 1$ and we shall henceforth restrict analysis to the range $1 \geq r \geq 0$, where $F_{R_0}(r)$ exhibits nontrivial behaviour. We then have $F_{R_0}(r) = 2\text{Prob}(|\xi_1| < r|\xi_2|) = 2\text{Prob}(r|\xi_2| > \xi_1 > -r|\xi_2|)$. It quickly follows that $F_{R_0}(r) = 2\int_{-\infty}^{\infty} f_\xi(y)[F_\xi(r|y|) - F_\xi(-r|y|)]\,\mathrm{d}y$. To obtain the probability density, we differentiate the above with respect to r and the result of the main text, Eq. (4), follows immediately.

To find the cumulative distribution and the probability density for Method 2, we simply note that if ξ_1, ξ_2, ξ_3 and ξ_4 are independently and identically distributed (i.i.d.), then $\zeta_1 = \xi_1 - \xi_2$ and $\zeta_2 = \xi_3 - \xi_4$ are also i.i.d. with zero mean, thus meeting the conditions of Method 1. Hence $F_R(r) = \text{Prob}(|\zeta_1|/|\zeta_2| < r \,|\, |\zeta_1| < |\zeta_2|)$ and the above results apply with the corresponding density and cumulative distribution of $\xi_i - \xi_j$. In this way we obtain Eqs. (6) and (7) of the main text.

Next we calculate the distributions of R_0 and R for some specific cases of interest. Because of the definitions of R_0 and R (Eqs. (2) and (5)), the cumulative distribution and probability density of these random variables are only non-zero for r in the range $1 \geq r \geq 0$ and for brevity, we shall only give the form of the distributions in this range of r.

A.1. Exponential distribution, Method 1

We calculate $F_{R_0}(r)$ by direct application of Eq. (3). We have $f_\xi(y) = \mathrm{e}^{-y}$ for $y \geq 0$ and zero otherwise. We also have $F_\xi(y) = (1 - \mathrm{e}^{-y})$ for $y \geq 0$ and zero otherwise. It follows that $F_{R_0}(r) = 2r/(1+r)$ and by differentiation or from Eq. (4) we obtain $f_{R_0}(r) = 2/(1+r)^2$.

A.2. Exponential distribution, Method 2

We have $f_\xi(y) = \mathrm{e}^{-y}$ for $y \geq 0$ and zero otherwise. We find $f_{\xi-\xi}(y) = \mathrm{e}^{-|y|}/2$ and application of Eqs. (6) and (7) leads to $F_R(r) = 2r/(1+r)$ and to $f_R(r) = 2/(1+r)^2$.

A.3. Reflected exponential distribution, Method 1

We have $f_\xi(y) = \mathrm{e}^{-|y|}/2$ and results for this distribution coincide with the results for the exponential distribution, for Method 2.

A.4. Reflected exponential distribution, Method 2

We have $f_\xi(y) = \mathrm{e}^{-|y|}/2$ and find $f_{\xi-\xi}(y) = (1 + |y|)\mathrm{e}^{-|y|}/4$. Application of Eqs. (6) and (7) lead to $F_R(r) = (r/2)(3 + 9r + 4r^2)/(1+r)^3$ and to $f_R(r) = (3/2)(1 + 4r + r^2)/(1+r)^4$.

A.5. Normal distribution, Method 1

We have $f_\xi(y) = \mathrm{e}^{-y^2/2}/\sqrt{2\pi}$ and application of Eqs. (3) and (4) lead to $F_{R_0}(r) = (4/\pi)\arctan(r)$ and $f_{R_0}(r) = 4/[\pi(1+r^2)]$.

A.6. Normal distribution, Method 2

We find that $f_{\xi-\xi}(y) = \mathrm{e}^{-y^2/4}/\sqrt{4\pi}$ and application of Eqs. (6) and (7) yield identical results, to those found for Method 1, for the cumulative distribution and probability density of R_0 for a normal distribution.

A.7. Uniform distribution, Method 1

Direct application of Eq. (3) for $\xi \sim U(0, 1)$ quickly yields $F_{R_0}(r) = r$ and hence $f_{R_0}(r) = 1$ (corresponding to $R_0 \sim U(0, 1)$).

A.8. Uniform distribution, Method 2

For $\xi \sim U(0, 1)$ we find $f_{\xi-\xi}(y) = (1 - |y|)$ for $1 > y > -1$ and is zero elsewhere. Direct application of Eq. (6) yields $F_R(r) = (r/3)(4 - r)$ and $f_R(r) = (2/3)(2 - r)$.

References

Aschoff, 1966. Circadian activity pattern with two peaks. Ecology 47, 662–667.

Ciofolo, I., Boissier, M., 1992. Diurnal fluctuations in activity in the lizard. J. Ethol. 10, 1–5.

Fisler, G.F., 1962. Variation in the morning awakening time of some birds in South-Central Michigan. Condor 64, 184–198.

Hogg, R.V., Tanis, E.A., 2006. Probability and Statistical Inference, 7th ed. Pearson.

Lilliefors, H.W., 1969. On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown. J. Am. Stat. Assoc. 64, 387–389.

Neave, H.R., 1989. Statistical Tables. Unwin Hyman.

Nouvellet, P., Bacon, J.P., Waxman, D., in preparation.

Semenov, Y., Ramousse, R., Le Berre, M., 2000. Effect of light and temperature on daily activities of the Alpine Marmot (Marmota marmota Linne, 1758) in its natural environment. Can. J. Zool.-Can. Zool. 78, 1980–1986.

Van Orsdol, K.G., 1984. Foraging behaviour and hunting success of lions in Queen Elisabeth National Park, Uganda. Afr. J. Ecol. 22, 79–99.

Weiss, N.A., 2006. A Course in Probability. Pearson.