Oecologia (1995) 103: 435–443

ORIGINAL PAPER

David R. Fox · James Ridsdill-Smith

Tests for density dependence revisited

Received: 26 October 1994 / Accepted: 13 April 1995

Abstract We have examined a number of statistical issues associated with methods for evaluating different tests of density dependence. The lack of definitive standards and benchmarks for conducting simulation studies makes it difficult to assess the performance of various tests. The biological researcher has a bewildering choice of statistical tests for testing density dependence and the list is growing. The most recent additions have been based on computationally intensive methods such as permutation tests and bootstrapping. We believe the computational effort and time involved will preclude their widespread adoption until: (1) these methods have been fully explored under a wide range of conditions and shown to be demonstrably superior than other, simpler methods, and (2) general purpose software is made available for performing the calculations. We have advocated the use of Bulmer's (first) test as a de facto standard for comparative studies on the grounds of its simplicity, applicability, and satisfactory performance under a variety of conditions. We show that, in terms of power, Bulmer's test is robust to certain departures from normality although, as noted by other authors, it is affected by temporal trends in the data. We are not convinced that the reported differences in power between Bulmer's test and the randomisation test of Pollard et al. (1987) justifies the adoption of the latter. Nor do we believe a compelling case has been established for the parametric bootstrap likelihood ratio test of Dennis and Taper (1994). Bulmer's test is essentially a test of the serial correlation in the (log) abundance data and is affected by the presence of autocorrelated errors. In such cases the test cannot distinguish between the autoregressive effect in the errors and a true density dependent effect in the time series data. We suspect other tests may be similarly affected, although this is an area for further research. We have also noted that in the presence of autocorrelation, the type I error rates can be substantially different from the assumed level of significance, implying that in such cases the test is based on a faulty significance region. We have indicated both qualitatively and quantitatively how autoregressive error terms can affect the power of Bulmer's test, although we suggest that more work is required in this area. These apparent inadequacies of Bulmer's test should not be interpreted as a failure of the statistical procedure since the test was not intended to be used with autocorrelated error terms.

Key words Population dynamics · Statistical tests · Simulation · Autocorrelation

D. R. Fox (✉)
IPP&P Biometrics Unit,
CSIRO Centre for Mediterranean Agricultural Research,
Private Bag, PO Wembley, WA, Australia 6014

J. Ridsdill-Smith
Division of Entomology,
CSIRO Centre for Mediterranean Agricultural Research,
Private Bag, PO Wembley, WA, Australia 6014

Introduction

The notion that biological populations have self-regulatory growth rates is one that has been fraught with difficulties (Pielou 1974; Dempster and Pollard 1986; Krebs 1991). The concept of density dependence has been a divisive issue among biologists, many regarding it as a mathematical abstraction rather than a biological reality. Fuelling the debate is the lack of a statistical paradigm for detection, estimation, and modelling of 'density-dependent' populations.

In a recent study, Holyoak (1993) compared the performance of a number of tests for density dependence and provided some broad recommendations on the basis of statistical power. In view of our own difficulty in trying to meaningfully compare different tests, we decided to review some of the important statistical issues relating to the detection of density dependence.

In doing so, we have drawn attention to some of the problems in assessing the utility of individual tests for density dependence. Our intent is to highlight the fact that the design of a definitive comparative study is perhaps as difficult as the construction of an 'optimal' test.

The development of a new test for density dependence is invariably motivated by a requirement to redress one or more shortcomings of an existing test or tests which have been identified in a previous comparative study. With each new test comes the inevitable performance comparisons and so the cycle repeats itself. Since these comparisons tend to yield 'fuzzy' conclusions, the unfortunate consequence of this process is that the biologist is being unwittingly encouraged to apply a battery of tests to a particular data set and to choose the most appealing result. Two difficulties are usually encountered:

1. The new test is not universally superior and so will have its own set of limitations.
2. Evaluation of the performance of the new test may depend in part on the methods used to construct the synthetic data on which the results are based. In particular, the representation of the model's stochastic component and the introduction of autocorrelation.

While it may not be possible to completely ameliorate these problems, we suggest that the adoption of some benchmarks for comparative studies would be useful. Furthermore, the use of a 'standard' test procedure against which all others could be compared would also be desirable. A suitable candidate is Bulmer's (first) test (Bulmer 1975).

Holyoak (1993) suggested that the randomisation test of Pollard et al. (1987) was the most reliable, while more recently Dennis and Taper (1994) have claimed up to a 50% improvement in power by using their parametric bootstrap likelihood ratio (PBLR) method. Both of these procedures have a heavy computational requirement which necessitates the use of a fast PC or workstation. In contrast, Bulmer's test statistic and critical values can be computed on a calculator with minimal effort. Furthermore, our own studies suggest that Bulmer's test performs well in a variety of situations and in a number of instances has better power than the PBLR and randomisation methods. Although Bulmer's test is not without its weaknesses and limitations, in terms of power "Bulmer's test is about the best we have" (Reddingius 1990).

The remainder of this paper is divided into two sections. The first deals with broad statistical issues associated with tests for density dependence and examines the strengths and weaknesses of Bulmer's test. The review of Holyoak's (1993) study in the second section serves to illustrate some of the difficulties with comparative type studies. Additional insight into problems of testing for density dependence is provided with some results of our own investigations.

Hypothesis tests and models

The early methods of Varley and Gradwell (1960) have been developed and refined in a number of ways so that practising biologists now have a formidable arsenal of statistical tests at their disposal. These include procedures due to Bulmer (1975), Pollard et al. (1987), Reddingius and den Boer (1989), Crowley (1992), Vickery and Nudds (1984), and most recently, Dennis and Taper (1994). Accompanying this relentless pursuit of the 'best' test has been a plethora of comparative studies and performance assessments (Maelzer 1970; St. Amant 1970; Itô 1972; Slade 1977; Bellows 1981; Gaston and Lawton 1987; Hassell et al. 1989; Solow 1990; Solow and Steele 1990; Vickery and Nudds 1991; Woiwod and Hanski 1992; Holyoak 1993; Wolda and Dennis 1993). Unfortunately, such studies generally fail to provide clear insights and recommendations that can be translated into standard statistical practice. Instead, we have a bewildering choice of statistical approaches with guidelines for their use along the lines of "test *a* is best for data conforming to model *b*, except when the effects of *c* are substantial, in which case test *d* is preferred". This may be satisfactory for simulation studies where the data are synthetic and generated from a *known* functional form; however, it is of limited practical value. There appears to be a lack of recognition of the fact that in simulation studies the *model* comes first and the *test* for density dependence second. In practice it is, or we believe should be, the other way around. So while it may be meaningful to talk about a test's 'specificity' in relation to simulated data this concept probably does not apply in practice. It is difficult to conceive that biological populations conform to some predetermined mathematical model such as an exponential logistic, a multiplicative logistic, or power model. Indeed, depending on parameter values chosen, each of these models may describe the observed population abundances equally well. We believe there is a burden of proof (Dennis and Taper 1994) that the case for density dependence must be established *before* models for its manifestations are contemplated. Furthermore, the test for density dependence should not be embedded in a particular model form. For example, one could test a null hypothesis of density independence (i.e. random walk in logarithms of population size) by examining H_0: $r = 0$ in the model $N_{t+1} = N_t \exp[r(1 - \alpha N_t)].\xi_t$ or by testing the hypothesis H_0: $\lambda = 1$; $a = 0$ in the model $N_{t+1} = \lambda N_t[1+(aN_t)^b]^{-1}.\xi_t$ (where ξ_t is a random error term). That different results arise because of the different functional forms and the manner in which parameters and their standard errors are calculated is often remarked upon. Holyoak and Lawton (1992) lamented the fact that different tests produced different results – an observation they ascribed to the different assumptions used in constructing the tests. Holyoak (1993) noted that "the

usefulness of individual tests will depend partly on which population model is the best descriptor of population data", while Dennis and Taper (1994) observed that conclusions among studies "seem to vary depending on which methods are used". It is also important to realise that differences between models arise not only from the use of different deterministic functions, but also by the way in which the stochastic component is represented for the *same* deterministic function. For example, both Holyoak (1993) and Dennis and Taper (1994) used a discrete logistic model of the form $N_{t+1} = N_t \exp[r(1-\alpha N_t)]$. The stochastic component in Holyoak's (1993) study was introduced by treating α as a random coefficient, while Dennis and Taper used fixed coefficients with an additive error component inside the exponent. Different results and possibly conclusions are to be expected and this aspect of comparative evaluations makes it particularly difficult for end-users to decide on a single test for use with their data.

We feel it is time to stand back and take a wider view of density dependence and associated statistical methodologies. First and foremost, the researcher should have a clear idea of the study objectives. For example, is one simply interested in establishing density dependent behaviour in a biological population or is there an additional requirement to *model* the population dynamics? If so, the question "what do we wish to do with the model?" needs to be answered. Procedures borrowed from time-series analysis may provide the researcher with good forecasting abilities but can fall short as a descriptive tool. The converse may be true for simple models such as the logistic. In his correspondence on Bulmer's (1975) paper, Anderson (1976) suggested using a Box-Jenkins approach (Box and Jenkins 1970) for modelling density dependence, although Bulmer (1976) remained unconvinced of the benefits of this approach. Bulmer's reluctance to use Box-Jenkins methods is understandable given that these techniques were relatively new and no doubt less well understood and used 20 years ago. We believe other 'contemporary' approaches to density dependent modelling (as distinct from testing for density dependence) should also be investigated. These might include, for example generalised linear models, generalised additive models, and Markov Chain Monte Carlo (MCMC) methods. Similar requests have been made by Wolda et al. (1994).

Bulmer's test

Ideally, a test of density dependence should satisfy the following criteria: (1) it makes minimal or no a priori assumptions about the response-generating mechanism; (2) it is simple to implement and interpret; and (3) it is unbiased and has desirable power characteristics for a wide range of models under the alternative hypothesis. Having applied the test to a particular data set the hypothesis of density *independence* is either accepted or rejected. The first outcome should signal that further modelling of density dependent behaviour is probably not warranted while the second outcome provides a prima facie case for more detailed modelling, estimation, and inference. Tests based on resampling procedures such as the randomisation test of Pollard et al. (1987) and the parametric bootstrap of Dennis and Taper (1994) are attractive and certainly fulfil conditions (1) and possibly (3) although these generally have a heavy computational requirement. Bulmer's test is a candidate, although it has been criticised by some authors for having poor power characteristics in the presence of temporal trends (Slade 1977; Pollard et al. 1987; Woiwod and Hanski 1992). It is common statistical practice to take remedial action when one or more assumptions of a parametric test have been violated. For example, violation of the homogeneity of variance assumption in analysis of variance can have disastrous effects on computed *P*-values (Horton 1978). Rather than discarding the method or seeking a less powerful non-parametric alternative, a variance-stabilising transformation of the data is often the first avenue of redress. Similarly, corrective procedures for Bulmer's test should be explored when temporal trends are present. Bulmer (1975) recommends against any attempts to 'detrend' the series prior to testing for density dependence. However, if the trend is an artefact of an atypical x_0 value, then one way to proceed might be to drop the first one or two observations from the series. This procedure is not uncommon in simulation studies where some small fraction of the initial simulated output is discarded to reduce 'start-up' effects. The decision to either accept or reject the null hypothesis of density independence should be independent of the choice of the initial value x_0, since this observation presumably corresponds to an arbitrary point in time. Values of x_0 which are far removed from the carrying capacity are most likely the result of atypical or extreme environmental factors and as such should not be incorporated into a test of density dependence. A possible test of the appropriateness of the first m observations, $\{x_0, x_1, \ldots, x_{m-1}\}$ $m \ll n$, for testing density dependence could be based on the change in some criterion (e.g. Bulmer's statistic) when these observations are omitted. Such 'leave-one-out' procedures are commonplace in regression (Belsley et al. 1980) and geostatistical analyses (Isaaks and Srivastava 1989) and we believe they have a role to play in testing for density dependence. More rigorous statistical investigations into the applicability and use of these 'cross-validation' techniques in tests for density dependence are required.

The Pollard et al. (1987) study of Bulmer's test was based on 200 simulations for series of length $n = 10$ generated from the model $x_{t+1} = r + \beta x_t + e_t$; $(\beta \neq 1)$, with fixed values of $r = 0.4$; $\beta = 0.8$, and normally

distributed errors having zero mean and variance (σ^2) of 0.01. A more comprehensive study would call for an examination of the effects due to changes in all five parameters (r, β, n, σ^2, x_0). For example, based on 1000 simulations with $r = 0.4$, $\beta = 0.3$, $n = 30$, $\sigma^2 = 0.1453$, and $x_0 = 0.5714$ we found Bulmer's test correctly rejected the hypothesis of density independence 99.2% of the time. However, when the initial value (x_0) was changed from the equilibrium value of 0.5714 to 0.25, the observed detection rate dropped to 53.8% and with $x_0 = 1.0$, the figure was 28.4%. We do not see this diminished performance as a failing of the test but rather a caveat to its use. Restrictive assumptions are a trait of most parametric tests, Bulmer's test is no different in this respect.

Finally, we note that Bulmer's test cannot discern whether the density dependence is direct or delayed although it has been shown (Holyoak 1994) that Bulmer's test will consistently reject the hypothesis of density independence when the data contain only delayed dependence.

Comparative studies

We have examined Holyoak's (1993) paper in an attempt to highlight some of the difficulties encountered in conducting comparative studies of various tests for density dependence. The randomisation test of Pollard et al. (1987) was endorsed by Holyoak as being the most 'reliable' (Holyoak 1993). However, the development of the parametric bootstrap test appears to have been motivated by a desire to redress the claimed low power of the Pollard et al. test (Dennis and Taper 1994). This is the most recent example of the comparative study cycle referred to in our introductory remarks.

The conclusions arising from Holyoak's investigation into the performance of some common tests of density dependence echo those made in a number of earlier studies (Itô 1970; Maelzer 1970; Holyoak and Lawton 1992; Woiwod and Hanski 1992 to name a few) and as such come as no great surprise. The recurring messages are: (1) inconsistent test conclusions frequently arise from different test procedures; (2) different test results are to be expected because of the different assumptions built into these tests; (3) rates of detection (power) are affected by the presence of autocorrelation, temporal trends, spatial trends, series length and assumed model form; (4) regression-based tests generally perform poorly; and (5) there is no such thing as a globally optimal test of density dependence.

Holyoak's study was based on an analysis of synthesised time series data 20 generations in length using a variety of models and parameter combinations. A mixture of fixed and random parameters were used for the density dependent models and the results of common tests for density dependence analysed using what the author terms a multivariate analysis of covariance. A number of important statistical considerations emerge and these are discussed in the following sections.

Choice of parameter values

Holyoak's choice of model parameters seems to have been fairly judicious so as to give a similar spread of abundances over the models considered. However, there were anomalies that were not fully explained (such as treatment of 'infeasible' parameter combinations and selection of initial values in the power model) which we found disquieting. Other problems associated with the selection of models and the assignment of parameter values have been identified by Wolda et al. (1994) and we shall not repeat them here. The point we wish to make is that the arbitrary way in which these difficulties are resolved represents a potential source of bias in the final outcome of comparative studies.

Representation of the stochastic component

While the random-coefficient models adopted by Holyoak are certainly a legitimate way of introducing variation into the population models, we nevertheless feel that this aspect of the study requires a more comprehensive investigation. In particular, it would be useful to know to what extent rates of detection are influenced by (1) non-normality of parameter and/or error distributions, and (2) correlated errors and/or parameter values. It is not our intention to conduct comprehensive simulation studies in this paper, but to report some preliminary investigations into these effects. The correlated errors issue is discussed in detail in the following section. Our assessment of the effects of non-normality is based on a limited simulation study using the so-called exponential logistic model (Holyoak's model 2) with values of the α parameter first generated from a normal distribution (as in Holyoak 1993) with mean 0.01 and standard deviation 0.0001 and secondly, from a displaced exponential distribution having the same mean and standard

Table 1 Rates of detection (based on 1000 simulations) for the exponential logistic model for selected n and r using normally distributed and exponentially distributed α values

r value	$n = 10$		$n = 20$		$n = 30$	
	Normal	Exponential	Normal	Exponential	Normal	Exponential
1.0	67.5	65.6	98.7	98.4	100.0	100.0
0.2	11.2	7.1	19.5	17.2	28.6	29.6
0.1	6.4	4.6	11.1	8.0	14.4	12.7

Table 2 Average rates of detection (%) for three tests of density dependence when applied to data generated from Dennis and Taper's (1994) logistic model with $b = -0.01$ (*PBLR* parametric bootstrap likelihood ratio)

Test method	Source	Average detection rate
Bulmer	Fox and Ridsdill-Smith	49.13
PBLR	Dennis and Taper (1994)	47.13
Randomisation (Pollard et al. 1987)	Dennis and Taper (1994)	36.53

deviation. The exponential distribution is highly skewed and thus represents a gross violation of the normality assumption.

One thousand series of length 10, 20, and 30 were generated for r values of 1.0, 0.2, and 0.1. The rates of detection for Bulmer's test for the normal and exponential cases are in reasonably close agreement with larger differences being observed for small series length and small r (Table 1).

In contrast, Dennis and Taper (1994) used the logistic model with fixed parameter values and a separate error term to compare the performance of their PBLR method with the Pollard et al. randomisation test. The results presented in their Table 7 (page 220) indicate the superiority of the PBLR method when applied to selected cases of density dependent data. We have reproduced the set of simulations for the density dependent data (corresponding to $b = -0.01$ in the Dennis and Taper model) and estimated the power of Bulmer's test using 1,000 simulations for each combination of other model parameters (n_0, σ, and a). Average rates of detection for all three methods for the $b = -0.01$ data are compared in Table 2.

On average, Bulmer's test had a marginally higher rate of detection than the PBLR method with the randomisation test performing worst. Interestingly, Holyoak found that the Pollard et al. test outperformed Bulmer's test for data generated from the logistic model (average rates of detection of 61.1% and 56.8% respectively). The point we wish to make is that neither study is definitive and the disparity of results is most likely due to the different treatments of the stochastic component and different parameter values used in each study.

Treatment of autocorrelation

Results presented by Holyoak were based on descriptive statistics of the synthesised observations and not on the properties of the parameter or error distributions. In particular, the autocorrelation referred to by Holyoak is in fact the (standardised) covariance between X_{t+1} and X_t which is not the same as the autocorrelation between the error terms.[1] As will be shown later, the former is a function of other model parameters and as such has no intrinsic interest. Autocorrelated errors are most likely to arise when the effect of the random disturbance is not instantaneous, but is likely to occur in future periods. It is difficult to speculate on the sources of autocorrelated errors, although it is possible that this phenomenon may be a consequence of environmental (both spatial and temporal) effects and/or measurement error.

Itô (1972) identified two sources of autocorrelation (serial correlation) in relation to the simple regression of a single dependent variable Y on a single independent variable x. The first of these involved a first-order, autoregressive [AR(1)] process for the assumed additive error terms and the second was what has become known in the econometric literature as a lagged dependent variable model (Fomby et al. 1988). By definition, density-dependent models represent a class of lagged dependent variable models. To take a simple case we assume the following density-dependent model:

$$x_{t+1} = \beta x_t + e_{t+1} \quad |\beta| < 1$$

where, without loss of generality, the x values have been centred about their mean. Assuming the errors are independently distributed with common variance σ_e^2, then the lag 1 covariance between the x values is $\beta\sigma_e^2/(1-\beta^2)$ and the first-order correlation coefficient is β. That Holyoak observed strong negative relationships between this autocorrelation and rates of detection is to be expected since the latter will increase as β (and hence the autocorrelation between the x values) becomes smaller. The true effects of autocorrelation can be examined through the model given above with an autoregressive structure for the error term, i.e.

$$e_{t+1} = \phi e_t + u_{t+1}$$

where the u_t are independently, identically distributed (iid) random variables with zero mean and variance σ_u^2. It is relatively easy to show that the covariance between the *regressor* x_t and error e_{t+1} for this model is

$$\frac{\phi\sigma_u^2}{(1-\phi^2)(1-\beta\phi)} \neq 0$$

This dependency between the regressor and the *current* error term is in contrast to the previous model where the regressors are related only to error terms of the *previous* periods. The effects of this 'contemporaneous' correlation (Fomby et al. 1988) on the power of Bulmer's test were examined by Reddingius (1990) in response to the claim by Solow (1990) that its presence

[1] Holyoak never intended to examine the effects of correlated errors (personal communication)

Table 3 Rates of detection (%) for Bulmer's test when applied to density dependent data generated using Eq. 1. Values for n refer to the series length. All results based on 1,000 simulations

β value	$\phi = -0.8$		$\phi = 0.0$		$\phi = 0.8$	
	$r = 0.2$	$r = 0.4$	$r = 0.2$	$r = 0.4$	$r = 0.2$	$r = 0.4$
$n = 10$						
−0.8	99.6	100	97.9	100	73.5	100
−0.3	99.3	98.9	83.3	85.1	26.9	27.7
0.3	93.5	92.9	40.2	41.1	3.2	2.6
0.8	72.3	72.5	8.4	10.0	0.4	0.4
$n = 20$						
−0.8	100	100	100	100	95.1	100
−0.3	100	100	100	99.9	51.2	51.6
0.3	100	100	87.1	86.5	4.2	3.9
0.8	95.2	95.2	19.1	18.8	0.4	0.3
$n = 30$						
−0.8	100	100	100	100	99.3	100
−0.3	100	100	100	100	67.1	69.9
0.3	100	100	98.7	98.7	5.9	6.5
0.8	99.8	99.7	28.2	29.6	0.0	0.0

did not affect the density independence in the null model when $\beta = 1$. Reddingius demonstrated that Solow had unfairly penalised Bulmer's test and dismissed Solow's claim that the test is non-robust and lacks power as being "not very helpful". Reddingius's rather casual remark that the introduction of autocorrelated errors is an "obvious 'small' modification of the original model" belies the potential impact of this effect.

In developing his test for density dependence, Bulmer (1975) believed β would lie in the range 0–1. Reddingius and den Boer (1989) attribute this to the fact that "for negative values of β, Bulmer's second test appears to be worthless" but suggested that there are no biological reasons for such a constraint. Our own simulation studies reveal that Bulmer's (first) test has exceedingly high power in such cases.

We have investigated the power of Bulmer's test for the lagged dependent variable model with autoregressive errors:

$$x_{t+1} = r + \beta x_t + e_{t+1} \quad |\beta| < 1$$
$$e_{t+1} = \phi e_t + u_{t+1} \quad |\phi| < 1 \qquad (1)$$

For $\beta = 1$ and $\phi = 0$ this is equivalent to a random walk ($r = 0$) or random walk with drift ($r \neq 0$) (Pollard et al. 1987). The mean of the x values changes with r and β while the variance is a function of σ_u^2, φ, and β. The rates of detection for Bulmer's test were obtained from 1,000 simulated data sets of length n for each of the parameter combinations considered (Table 3). The initial value, x_0, was set equal to the equilibrium value and σ_u^2 was adjusted to give an approximate 8% coefficient of variation for the x values.

A number of features are evident from Table 3: (1) generally speaking, rates of detection increase with increasing series length; (2) there is a negligible effect between different r values used – this is attributed to the 'standardisation' of initial values, series mean and variance as a function of r; (3) the rate of detection is almost 100% when φ is negative – irrespective of the value of β; (4) rates of detection are higher for negative β values; and (5) in the absence of autocorrelation ($\phi = 0$) the rate of detection for positive β values is acceptable only for series of at least length 30. Positive β and/or φ values require longer series for the density dependence to be detected.

There is nothing particularly new in these findings although the most interesting phenomenon, which has not been previously studied in great detail, is the effect of autoregressive errors. As pointed out by Holyoak and others, Bulmer's test is essentially a test of autocorrelation (one only need compare Bulmer's statistic with the Durbin-Watson test statistic for serial correlation to see this). Thus, it is difficult to conclude in any particular instance whether or not rejection of a null hypothesis is due to a density dependent effect, autocorrelated errors (which can arise from external, environmental impacts), or a combination of both.

For the density *independent* case the null hypothesis has been incorrectly rejected about 5% of the time when the error terms are independent ($\phi = 0$) and there is no drift in the series ($r = 0$) (Table 4). In the presence of drift, Bulmer's test becomes conservative in the extreme with no rejection of the null hypothesis. More importantly is the test's sensitivity to negatively autocorrelated errors ($\phi = -0.8$). The crucial observation is that in the presence of autocorrelated errors, the type I error rate is nowhere near the nominal 5%. We conclude that, in such instances, the detection rate is based on a faulty significance region and therefore comparisons between tests in the presence of autocorrelated errors are rather meaningless.

This situation parallels closely the general requirement in statistical estimation for *unbiasedness* when assessing the quality of competing estimators. The notion of unbiasedness is also used in hypothesis testing. Loosely speaking, a statistical test is unbiased if the maximum power of the test (where power is defined as the probability that the test rejects the null hypothesis) when the null hypothesis is *true*, does not exceed the level of significance. Clearly, Bulmer's test (as it stands) is not unbiased in the presence of autocorrelated errors. Furthermore, the results presented in

Table 4 Rates of detection (%) for Bulmer's test (5% level of significance) using random walk data (density independent) generated using Eq. 1 with $\beta = 1$. All results based on 1,000 simulations

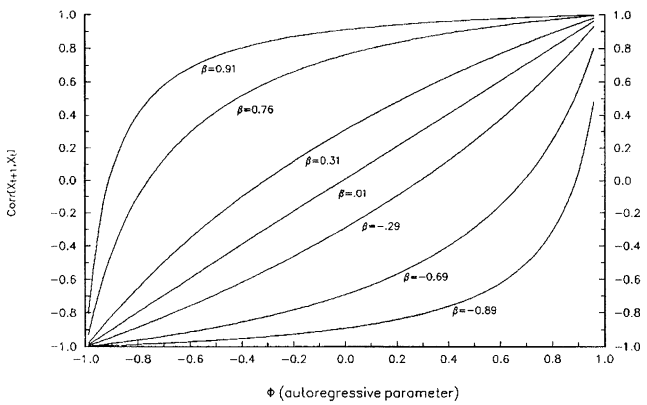
n	$\phi = -0.8$		$\phi = 0.0$		$\phi = 0.8$	
	$r = 0$	$r = 0.4$	$r = 0$	$r = 0.4$	$r = 0$	$r = 0.4$
10	55.1	0.0	4.7	0.0	0.3	0.0
20	72.0	0.0	6.5	0.0	0.1	0.0
30	76.9	0.0	6.1	0.0	0.0	0.0

Holyoak's Table 1 would suggest that the regression procedures and possibly Crowley's test are biased even in the *absence* of autocorrelated errors. Thus it is doubtful as to whether power comparisons involving these tests will permit any meaningful conclusions.

In defense of Bulmer's and other parametric tests, we repeat our earlier sentiment that this should not be construed as a failing of the test, but rather an indicator that special care must be taken to identify and take account of violations of a test's assumptions. In this respect we add our support to Reddingius' (1990) claim that Bulmer's test has been unfairly penalised since we are requiring it to perform in situations for which it was never intended to be applied. Bulmer's test was developed under the explicit assumption that $\phi = 0$. When this is the case the two sets of hypotheses $\{H_0: \rho = 1$ versus $H_1: \rho < 1\}$ and $\{H_0: \beta = 1$ versus $H_1: \beta < 1\}$ are equivalent and no difficulties arise. However, when $\phi \neq 0$, hypotheses couched in terms of ρ correspond to an infinite number of β values, depending on the value of ϕ, and so the results are unreliable. To illustrate, consider two variations of Eq. (1): suppose model I has parameters $\{r = 0;\ \beta = 0.5;\ \phi = -0.8\}$ while another model, model II, has parameters $\{r = 0;\ \beta = -0.8;\ \phi = 0.5\}$. In both models I and II the first-order serial correlation between the x values is $\rho = -0.5$, and therefore Bulmer's test will reject the hypothesis of density independence with the same power even though the density dependent effect in model II is far more pronounced. This indeterminacy is succinctly revealed by writing both parts to Eq. 1 as a single model using the backward shift operator, B where $BX_t = X_{t-1}$ (Box and Jenkins 1970): $(1-\phi B)(1-\beta B)x_t = u_t$. Using this representation it is apparent that ϕ and β are completely exchangeable. Finally, we note that the same model is produced by setting either $\phi = 0$ or $\beta = 0$ in Eq. 1. So if ϕ has to be estimated we expect there to be little additional information about β and therefore *any* test of density dependence which allows for $\phi \neq 0$ will have low power.

It can be established that for the autoregressive model just considered, the first-order serial correlation between the x values is $(\phi + \beta)/(1 + \beta\phi)$ (note that when $\phi = 0$ this reduces to β as before). The *inverse* relationship between a test's power and β has already been noted for the $\phi = 0$ case. Clearly, this effect can be amplified by making ϕ *negative*. Thus as $\phi \to -1$, so too does the serial correlation between the x values – regardless of the value of β. This is graphically illustrated in Fig. 1 where the serial correlation has been plotted as a function of ϕ for selected values of β. Thus, Holyoak's empirically based conjecture that more positive autocorrelation (first-order serial correlation between the x values) "would be expected to decrease the detection rates of those tests which rely on it to show the presence of density dependence" is correct.

Fig. 1 Serial correlation between x values as a function of density dependent parameter (β) and autoregressive parameter (ϕ)



Fixed series length

A major limiting factor in many density dependent studies is the small record length of the biological data and the corresponding low power of the statistical tests. For the biologist, $n = 10$ represents a long series when the data are annual census counts. However, from a statistical point of view, inference based on samples of size ten is treacherous. Reddingius (1990) remarked that $n = 25$ is not a statistically large sample size and so "we cannot have very large power of our tests". The comparisons presented in Pollard et al. (1987) were all based on $n = 10$. This arose not so much from any biological considerations as computational ones – there being $9! = 362{,}880$ permutations of the x_i values to be considered. The permutation tests may afford some advantage over parametric counterparts for small values of n, although our feeling is that tests of density dependence are really only reliable for series of at least 20 and preferably 30 or more.

Number of simulation runs

The seemingly large number of simulations (e.g. 3,600) reported in Holyoak's Table 1 is not as impressive as this figure suggests. A reasonable number of parameter combinations have been represented and for each of these, 25 simulated series were constructed so that for any model/parameter combination the resolution of detection rates is only 0.04. The large number of simulation runs reported comes from pooling the individual detection rates over all parameter combinations. We suggest a minimum of 100 simulation runs for each parameter combination would have been a more appropriate number. It is interesting to note that, in another comparative study by the same author (Holyoak 1994), 100 simulations were used and that an initial 2000 generations were discarded to avoid the start-up effects referred to earlier in this paper.

Statistical analysis of results

Holyoak has identified "statistics which bias detection of density dependence" by analysing the results of his simulation study using what he termed a "multivariate analysis of covariance" with the GLIM software. The analysis is in fact equivalent to a series of *univariate* logistic regressions (the GLIM software has no intrinsic multivariate capabilities) – there being one for each of the seven test procedures considered. The dependent variable in each case is the rate of detection; the 'explanatory' variables are statistics derived from the simulated series and include: first-order serial correlations, range, trend, skewness, kurtosis, mean, variance and a variance : mean ratio; a logit link was used to relate the mean rate of detection to these variables; the assumed error distribution was binomial. We have reservations about the utility of this analysis. Firstly, the covariates are themselves all artefacts of the model used to generate the data. Surely it would make more sense to attempt to relate rates of detection to one or two model parameters rather than a host of summary statistics which themselves are derived from the model and in some instances highly correlated (e.g. range and variance; mean, variance and mean : variance ratio). More importantly, the lack of independence between rates of detection for the various model forms has been ignored in Holyoak's analysis. The effects of this on the reported P values could be quite serious. There also needs to be some justification for the choice of model used to relate rates of detection to potential explanatory factors. The form chosen is common for the analysis of proportions, but so too are other formulations using different error/link combinations.

Conclusions

The paper by Holyoak (1993) has been examined in some detail since it seems to typify the current status of performance evaluation for tests of density dependence. Overall, Holyoak's study failed to clarify important issues relating to the identification of density dependence in biological populations. There were caveats ("all findings of this sort must remain tentative ...") and uncertainties ("a number of tests appear to be tests of autocorrelation") which weakened any impact of the results and diminished our confidence in the conclusions. Our uneasiness was also heightened by the small number of simulations, the absence of any investigation into the effects of varying series length, non-normality and lack of independence in parameter/error distributions. Furthermore, we are not convinced that the GLIM analysis has been the most meaningful way to interpret the simulation results, or the most appropriate from a statistical point of view. The important issue of autocorrelation has been dealt with only superficially, and has focused on the first-order serial correlation between log-abundances and not on what we consider to be the more relevant lack of independence in either error terms or parameter values. Additional problems with Holyoak's (1993) study have been identified in Wolda et al. (1994).

Biologists have certainly benefited from the tools that mathematicians and statisticians have been able to provide for analysing complex dynamical systems. However, in the case of density dependence, they have also been confused by the bewildering array of options at their disposal. In this paper we have sought to highlight some of the difficulties associated with procedures for test development and evaluation, and the problems of 'fuzzy' conclusions. We have advocated the adoption of Bulmer's (first) test as a de facto standard on the grounds that it makes minimal a priori assumptions about the response-generating mechanism, it has reasonable power characteristics over a wide range of alternative hypotheses and is easy to implement. We have drawn attention to some of this test's weaknesses, and indicated possible avenues for further investigation.

Acknowledgements We gratefully acknowledge the helpful suggestions of Barry Longstaff and Paul Wellings, CSIRO Division of Entomology (IPPP) and thank Richard Morton, CSIRO Biometrics Unit (INRE), for identifying a number of important statistical issues. Bert deBoer provided assistance with the simulations. We thank Marcel Holyoak for his careful review of an earlier draft of the manuscript and for clarifying a number of important points.

References

Anderson OD (1976) On Bulmer's statistical analysis of density dependence. Biometrics 32: 485–486

Bellows TS (1981) The descriptive properties of some models for density dependence. J Anim Ecol 50: 139–156

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York

Box GEP, Jenkins GM (1970) Time series analysis: forecasting and control. Holden-Day, California

Bulmer MG (1975) The statistical analysis of density dependence. Biometrics 31: 901–911

Bulmer MG (1976) Reply to Anderson (1976). Biometrics 32: 486–487

Crowley PH (1992) Density dependence, boundedness, and attraction: detecting stability in stochastic systems. Oecologia 90: 246–254

Dempster JP, Pollard E (1986) Spatial heterogeneity, stochasticity and the detection of density dependence in animal populations. Oikos 46: 413–416

Dennis B, Taper ML (1994) Density dependence in time series observations of natural populations: estimation and testing. Ecol Monogr 64(2): 205–224

Fomby TB, Hill RC, Johnson SR (1988) Advanced econometric methods. Springer, Berlin Heidelberg New York

Gaston KJ, Lawton JH (1987) A test of statistical techniques for detecting density dependence in sequential censuses of animal populations. Oecologia 74: 404–410

Hassell MP, Latto J, May RM (1989) Seeing the wood for the trees: detecting density-dependence from existing life-table studies. J Anim Ecol 58: 883–892

Holyoak M (1993) New insights into testing for density dependence. Oecologia 93: 435–444

Holyoak M (1994) Identifying delayed density dependence in time-series data. Oikos 70: 296–304
Holyoak M, Lawton JH (1992) Detection of density dependence from annual censuses of bracken-feeding insects. Oecologia 91: 425–430
Horton RL (1978) The general linear model. McGraw-Hill, New York
Isaaks EH, Srivastava RM (1989) Applied geostatistics. Oxford University Press, Oxford
Itô Y (1972) On the methods for determining density-dependence by means of regression. Oecologia 10: 347–372
Krebs CJ (1991) The experimental paradigm and long-term population studies. Ibis 133: 3–8
Maelzer DA (1970) The regression of log N_{n+1} on log N_n as a test of density dependence: an exercise with computer-constructed density-dependent populations. Ecology 51: 810–822
Pielou EC (1974) Population and community ecology. Principles and methods. Gordon and Breach, New York
Pollard E, Lakhani KH, Rothery P (1987) The detection of density dependence from a series of annual censuses. Ecology 68: 2046–2055
Reddingius J (1990) Models for testing. A secondary note. Oecologia 83: 50–52
Reddingius J, Boer PJ den (1989) On the stabilization of animal numbers. Problems of testing. 1. Power estimates and estimation errors. Oecologia 78: 1–8
Slade NA (1977) Statistical detection of density dependence from a series of sequential censuses. Ecology 58: 1094–1102
Solow AR (1990) Testing for density dependence: a cautionary note. Oecologia 83: 47–49
Solow AR, Steele JH (1990) On sample size, statistical power, and the detection of density dependence. J Anim Ecol 59: 1073–1076
St. Amant JLS (1970) The detection of regulation in animal populations. Ecology 51: 823–828
Varley GC, Gradwell GR (1960) Key factors in population studies. J Anim Ecol 29: 399–401
Vickery WL, Nudds TD (1984) Detection of density dependent effects in annual duck censuses. Ecology 65: 96–104
Vickery WL, Nudds TD (1991) Testing for density-dependent effects in sequential censuses. Oecologia 85: 419–423
Wolda H, Dennis B (1993) Density dependent tests, are they? Oecologia 95: 581–591
Wolda H, Dennis B, Taper ML (1994) Density dependent tests, and largely futile comments: answers to Holyoak and Lawton (1993) and Hanski, Woiwod and Perry (1993). Oecologia 98: 229–234
Woiwod IP, Hanski I (1992) Patterns of density dependence in moths and aphids. J Anim Ecol 61: 619–629