Limnol. Oceanogr., 45(2), 2000, 506–509

Comparing two vertical plankton distributions

Abstract—Traditional statistical methods for testing for differences in vertical plankton distributions are invalid when the distributions are patchy. This note describes a modification of the Kolmogorov-Smirnov test that is insensitive to patchiness. The modified test is applied to some data on day–night differences in the vertical distribution of two zooplankton taxa.

Vertical heterogeneity of physical, chemical, and biological properties is one of the most striking features of the world's lakes and oceans (Sverdrup et al. 1942; Hutchinson 1967). From a biological perspective, where and when an organism is located in the vertical dimension has important implications at the individual, population, and community levels (Russell 1927; Cushing 1951; Banse 1964; Longhurst 1976; Lampert 1989). For example, in the case of zooplankton, vertical location affects temperature-dependent metabolic processes, food availability, and predation risk. In conjunction with other factors, these determine the growth and survival of individuals. The integration of these vital rates determines population characteristics such as fecundity, stage duration, and mortality; and these, in turn, determine the absolute and relative abundances of populations comprising a zooplankton community.

In undertaking field and laboratory studies of vertical plankton distributions, it is often necessary to compare in a statistical way distributions under different conditions. Venrick (1986) pointed out that patchiness in these distributions invalidates traditional statistical methods for comparing them. This note revisits this problem in a slightly more formal setting and describes a method that can be used when traditional methods fail.

Suppose that we wish to compare the vertical distributions of two populations of plankton. For example, these populations could refer to groups of individuals at the same sampling station at different points of the diel cycle. Let π_{jk} be the probability that an individual selected at random from population j ($j = 1, 2$) occupies depth bin k ($k = 1, 2, \ldots, K$). Interest centers on testing the null hypothesis that the vertical distributions are the same for the two populations. This can be expressed formally as H_0: $\pi_{jk} = \pi_k$ for all k. To test H_0, each population is sampled in each of K depth bins. Let n_{jk} be the number of individuals of population j observed in depth bin k. It is convenient to arrange these counts in a 2-by-K table

$$n_{11}n_{12} \ldots n_{1K}$$

$$n_{21}n_{22} \ldots n_{2K}$$

A common statistic used to test H_0 is the Kolmogorov-Smirnov statistic, defined as

$$D = \max(|P_{1k} - P_{2k}|),$$

where

$$P_{jk} = \sum_{l=1}^{k} p_{jl}$$

is the sample cumulative distribution function of depth for population j and where p_{jk} is the usual estimate of π_{jk}

$$p_{jk} = n_{jk}/n_j,$$

where

$$n_j = \sum_{k=1}^{K} n_{jk}$$

The Kolmogorov-Smirnov statistic is known to have high power against so-called shift alternatives (Darling 1957). Two distributions differ by a shift if it is possible to transform one into the other by a unidirectional reallocation of probability mass (i.e., by individuals moving only up or only down).

To assess the significance of the observed value of D, it is necessary to know its sampling distribution under H_0. To begin with, suppose that individuals are distributed independently. Although standard results on the distribution of D are not valid for binned data (Noether 1963), significance can be assessed by the following randomization procedure. Pool the two samples into a single sample of $n_1 + n_2$ individuals. Divide the pooled sample at random into two samples of sizes n_1 and n_2. Calculate the value of D for these two samples. Repeat the procedure a large number of times and estimate the significance level (or P-value) by the proportion of simulated values of D that exceed the observed value. This is an example of a randomization test (Manly 1991). This is the same randomization scheme underlying Fisher's exact test of independence in a contingency table. The test is not sensitive to differences in overall abundance between the two samples but only to differences in the distribution of relative abundance with depth.

As Venrick (1986) pointed out, the assumption that the individuals within each sample are distributed independently of each other is critical to the validity of the Kolmogorov-Smirnov test. This also applies to the randomization test outlined above. To understand in qualitative terms the problem introduced by dependence, it is instructive to consider an

Table 1. Estimated rates of false rejection of H_0 for testing at the nominal 0.05 significance level for methods (1) and (2) of estimating the significance level for the three test statistics D, W, and W' and for three values of the variance : mean ratio γ.

	$\gamma = 1$		$\gamma = 5$		$\gamma = 10$	
	(1)	(2)	(1)	(2)	(1)	(2)
D	0.048	0.054	0.628	0.644	0.824	0.830
W	0.026	0.086	0.006	0.110	0.004	0.096
W'	0.042	0.042	0.058	0.058	0.046	0.046

Table 2. Estimated power for testing H_0 at the nominal 0.05 significance level for method (2) of estimating the significance level for the two test statistics W and W' and for three values of the variance : mean ratio γ.

	$\gamma = 1$	$\gamma = 5$	$\gamma = 10$
W	0.774	0.474	0.344
W'	0.728	0.466	0.340

extreme situation. Suppose that, in each sample, individuals arrive sequentially. The probability that the first individual in each sample chooses depth bin k is π_k, so that H_0 is true. For behavioral reasons, further individuals choose the depth bin in which the density of previously arrived individuals is greatest. This will result in extreme patchiness; in fact, all individuals in each sample will occupy the same depth bin. Suppose that the first individuals in the samples choose different depth bins. The corresponding value of D will be equal to 1, its maximum possible value. However, the randomization scheme outlined above will give rise to distributions in which two depth bins are occupied in each sample. Such distributions, which cannot arise under this extreme form of patchiness, have values of D less than 1, making the observed value of D appear to be extreme. By giving rise to values of D that cannot actually occur, the randomization test will provide an incorrect assessment of significance. Although this situation is extreme, it illustrates the way in which patchiness alone can lead to differences in vertical distributions that are more extreme than expected under the independence assumption.

Unfortunately, it is not possible to distinguish between patchiness and true differences in vertical distribution on the basis of unreplicated samples. Were replicate samples available, the observed variability at the same depth between replicates within samples would provide information about patchiness. For example, in the absence of patchiness, the replicate counts would exhibit Poisson variability (i.e., a variance-to-mean ratio of 1), whereas patchiness would result in overdispersion (i.e., a variance-to-mean ratio greater than 1). Not only would replicate sampling allow the detection of patchiness, but it would provide access to a range of methods for testing H_0. In this note, attention is restricted to the case of unreplicated sampling, and the case of replication is taken up elsewhere.

Smith et al. (1998) described a modification of D whose sampling distribution under H_0 is insensitive to overdispersion. The modified statistic is given by $W = D/L$ where

$$L = \sum_{k=1}^{K} |p_{1k} - p_{2k}|$$

In further work, we have found that the discreteness of the distribution of W can pose a problem, particularly when K is less than around 8, as is often the case in practice. Specifically, if the significance level is estimated by the proportion of randomizations for which the value of W is strictly greater than the observed value of W, then the true rate of false rejection of H_0 is less than the nominal significance level and the test has low power. On the other hand, if the significance level is estimated by the proportion of randomizations for which the value of W is less than or equal to the observed value of W, then the power of the test is higher, but the true rate of false rejection exceeds the nominal significance level. We have also found that replacing L by the closely related quantity

$$L' = \left(\sum_{k=1}^{K} (p_{1k} - p_{2k})^2\right)^{1/2}$$

improves the performance of the test, in the sense that the true rate of false rejection of H_0 is close to the nominal rate without a sacrifice of power.

To illustrate this, we present some results extracted from a larger simulation study. Suppose that $K = 6$, and, to begin within, suppose that the bin counts have a negative binomial distribution with mean

$$\mu_{jk} = \mu_k = 34 - 4k$$

and variance : mean ratio γ. The negative binomial distribu-

Table 3. Individual counts of two zooplankton species in six depth bins for two daytime and two nighttime samples.

Depth bin	Depth range (m)	Day		Night	
		Sample 1	Sample 2	Sample 1	Sample 2
Calanus pacificus					
1	0–10	0	1	121	120
2	10–25	7	2	24	108
3	25–50	6	3	16	32
4	50–75	25	44	7	1
5	75–125	12	50	9	1
6	125–175	12	3	0	0
Metridia lucens					
1	0–10	1	0	63	100
2	10–25	52	11	1092	49
3	25–50	8	2	137	8
4	50–75	6	3	10	7
5	75–125	1824	449	30	10
6	125–175	828	240	68	8

Table 4. Estimated significance levels for the test based on W' for comparing two daytime samples and two nighttime samples for the data given in Table 3. Significance levels were estimated from 1,000 randomizations.

	Calanus pacificus	*Metridia lucens*
Day 1 versus day 2	0.840	0.525
Night 1 versus night 2	0.615	0.450

Table 5. Estimated significance levels for the test based on W' for daytime–nighttime comparisons for the data in Table 3. Significance levels were estimated from 1,000 randomizations.

	Calanus pacificus	*Metridia lucens*
Day 1 versus night 1	0.220	0.140
Day 1 versus night 2	0.020	0.045
Day 2 versus night 1	0.110	0.205
Day 2 versus night 2	0.005	0.055

tion is the standard model for overdispersed counts. If $\gamma = 1$, it corresponds to the Poisson distribution. We simulated 500 sets of counts according to this model for $\gamma = 1, 5, 10$ and tested H_0 at the nominal 0.05 significance level using D, W, and the modified test statistic $W' = D/L'$. For each set of simulated counts, the tests were based on 500 randomizations. Table 1 gives the proportion of simulations for which H_0 was rejected using each test statistic when the significance level was estimated by each of two methods: (1) as the proportion of randomizations for which the value of the test statistic is strictly greater than its observed value and (2) as the proportion of randomizations for which the value of the test statistic is greater than or equal to its observed value. As H_0 is true for this experiment, for the test to be valid, the entries in Table 1 should be close to 0.05.

Turning to Table 1, the poor performance of D under overdispersion is clear: H_0 is rejected incorrectly at a rate much higher than the nominal significance level. The rejection rate of H_0 using W is well below 0.05 using method (1) and well above 0.05 using method (2). For W', the rejection rate of H_0 is close to 0.05 using either method.

Suppose, now, that the bin counts have mean

$$\mu_{1k} = 34 - 4k \qquad \mu_{2k} = 6 + 4k.$$

Table 2 gives the proportion of simulations for which H_0 was rejected when method (2) was used for estimating the significance level. In this case, H_0 is false and the entries in Table 2 are estimates of the power of the tests. As the significance levels for the test based on D are much higher than the nominal 0.05 level for $\gamma > 1$, results for D are not presented. Also, as the power of W was very low when method (1) was used to estimate the significance level, only results for method (2) are presented. Finally, to ensure that the power comparison is fair, the test based on W was performed at the nominal 0.05 significance, but the test based on W' was performed at the nominal 0.10 significance level, which is close to the true rejection rate for W given in Table 1. The results in Table 2 show that the better behavior of the test based on W' shown in Table 1 does not come at the cost of reduced power.

As noted, these results were extracted from a larger simulation study. Although they are fairly typical, the performance of W is considerably worse when K is smaller, the mean bin counts are lower, or γ is greater. In contrast, W' continues to perform well in such situations. As the test based on W' is valid, in the sense that the actual and nominal significance levels agree, and as this validity involves little or no sacrifice in power or computational simplicity, this statistic is preferable to W.

We conclude by presenting results from applying the test based on W' to some data concerning the vertical distribution of *Calanus pacificus* and *Metridia lucens*. These data, which are given in Table 3, were extracted from Bollens et al. (1993), and the original paper should be consulted for details. For each species, the data consist of two samples collected during the day and two samples collected during the night at each of $K = 6$ depth bins. Table 4 presents the results of comparing daytime samples with daytime samples and nighttime samples with nighttime samples, and Table 5 presents the results of comparing daytime samples with nighttime samples. In each case, the significance level was estimated from 1,000 randomizations. As the true significance levels for the tests based on D and W are unknown, we do not present results for these tests. The results in Table 4 indicate no evidence of differences between the two daytime samples or between the two nighttime samples for either species. In contrast, there is clear evidence in Table 5 of diel variation for each species. It is interesting to note, however, that for both species the comparisons involving night 1 are less significant than those involving night 2. Some of the significance levels in Table 5 may seem rather high. This is a reflection of the fact that the extremely high counts in some bins may be attributable to patchiness.

Andrew R. Solow[1]

Woods Hole Oceanographic Institution
Woods Hole, Massachusetts 02543

Stephen M. Bollens

Biology Department and Romberg Tiburon Center
San Francisco State University
P.O. Box 855
Tiburon, California 94920

Andrew Beet

Woods Hole Oceanographic Institution
Woods Hole, Massachusetts 02543

References

BANSE, K. 1964. On the vertical distribution of zooplankton in the sea. Prog. Oceanogr. **2:** 53–125.

BOLLENS, S. M., B. W. FROST, K. OSGOOD, AND S. D. WATTS. 1993.

[1] Corresponding author (asolow@whoi.edu).

Acknowledgements

The helpful comments of two anonymous reviewers are acknowledged with gratitude.

Vertical distributions and susceptibilities to vertebrate predation of the marine copepods *Metridia lucens* and *Calanus pacificus.* Limnol. Oceanogr. **38:** 1841–1851.
CUSHING, D. H. 1951. The vertical migration of planktonic crustacea. Biol. Rev. **26:** 158–192.
DARLING, D. A. 1957. The Kolmogorov-Smirnov, Cramer-von Mises tests. Annals of Mathematical Statistics **28:** 823–838.
HUTCHINSON, G. E. 1967. A treatise on limnology. Vol. II. Wiley.
LAMPERT, W. H. 1989. The adaptive significance of diel vertical migration of zooplankton. Funct. Ecol. **3:** 21–27.
LONGHURST, A. R. 1976. Vertical migration, p. 116–137. *In* D. Cushing and J. Walsh [eds.], The ecology of the seas. Blackwell.
MANLY, B. F. 1991. Randomization and Monte Carlo tests in biology. Chapman and Hall.
NOETHER, G. E. 1963. Note on the Kolmogorov-Smirnov statistic in the discrete case. Metrika **7:** 115–116.
RUSSELL, F. S. 1927. The vertical distribution of plankton in the sea. Biol. Rev. **2:** 213–262.
SMITH, W., A. BEET, AND A. R. SOLOW. 1998. Testing for shifts in the vertical distribution of plankton using a robust Kolmogorov-Smirnov-like statistic. J. Agric. Biol. Environ. Stat. **3:** 421–429.
SVERDRUP, S. V., M. V. JOHNSON, AND R. H. FLEMING. 1942. The oceans: Their physics, chemistry, and general biology. Prentice-Hall.
VENRICK, E. L. 1986. The Smirnov statistic: An incorrect test for vertical distribution patterns. Deep-Sea Res. **33:** 1275–1277.

Received: 31 October 1998
Amended: 1 October 1999
Accepted: 9 November 1999

Limnol. Oceanogr., 45(2), 2000, 509–515

Sequential extraction and analysis of phosphorus in marine sediments: Streamlining of the SEDEX procedure

Abstract—We streamlined the five-step SEDEX (sedimentary extraction) procedure for characterizing sedimentary phosphorus to a four-step procedure. We combined extraction of adsorbed and oxide-associated pools into a single step, retaining steps for authigenic, detrital, and organic P. In addition, we used automated spectrophotometric flow injection analysis (FIA) to determine P concentrations, rather than traditional spectrophotometric techniques. We decreased our total extraction and analytical time from 12 d for 24 replicate samples to 5 d without sacrificing our geochemical objectives, detection limits, or analytical reproducibility.

The oceanic history of reactive phosphorus (P) (dissolved P available to promote oceanic primary productivity) is of interest because of the role of P as a biolimiting nutrient. Thus, knowledge of reactive P burial in marine sediments is a key to testing hypotheses about temporal changes in P input or output fluxes. In addition, geochemical understanding of P transformations provides critical insights into the diagenetic behavior of P that controls its retention in sediments. Thus, we are interested in distinguishing the reactive from the nonreactive P components in sedimentary burial and, within the reactive fraction, the authigenic from the organic and oxide-sorbed P fractions.

The SEDEX (sedimentary extraction) procedure operationally defines five P components (adsorbed, oxide-associated, authigenic, detrital, and organic), and was extensively tested for use in marine sediments with analogs for naturally occurring phosphatic phases (Ruttenberg 1992). The procedure allowed critical insights into the transition of organic, adsorbed, and oxide-associated P to authigenic P (Sink-swapping, Ruttenberg and Berner 1993). SEDEX has now been applied to a variety of oceanic sediments, adding evidence for P sink-swapping in deep ocean sediments and allowing the quantification of reactive P burial (e.g., Filippelli and Delaney 1996; Delaney and Anderson 1997).

In this paper, we propose the removal of the operationally defined adsorbed step from the SEDEX procedure (renamed, four-step procedure). It is important to note that analogue phases are not available for the adsorbed and oxide-associated steps as sorptive surfaces are not pure phases (Ruttenberg 1992; Coston et al. 1995). We deleted the adsorption step on the assumption that most sorbed (combination of adsorbed and coprecipitated) phosphate associates with aluminum, iron, and mixed oxides (e.g., Turner et al. 1981; Coston et al. 1995; Slomp et al. 1996). Although there will be some adsorptive sites available on clays, the alkaline character and high ionic strength of seawater and interstitial waters means that few positive sorptive sites are likely to be available (e.g., Sposito 1984). Thermodynamic gains are typically much greater in the sorption of phosphate on oxides than they are on clays, and thus, as long as oxide surfaces are available, P will sorb to them (e.g., Sposito 1984). Furthermore, with the advent of techniques that allow observations on the binding characteristics of elements on surfaces, investigators have been able to observe how oxyanions such as P bind to metal surfaces. In work on the chemically analogous oxyanion arsenate, investigators have found that arsenate is not incorporated into the crystalline structure of oxide precipitates, but rather is adsorbed on the internal surfaces of the less organized oxide precipitates. As the oxides become more crystalline, arsenate is bumped out of the crystal structure (Fuller et al. 1993; Waychunas et al. 1993). Similar trends have been noted with phosphate (Fuller pers. comm.). Thus, the distinction between adsorbed and oxide-associated P in the SEDEX procedure may be in and of itself rather arbitrary, as most of the phosphate may be adsorbed on oxides. Thus, the four-step procedure may better define an operational sorbed fraction.