Research

Author for correspondence:
Graeme D. Ruxton
e-mail: graeme.ruxton@st-andrews.ac.uk

Testing for departure from uniformity and estimating mean direction for circular data

Graeme D. Ruxton

School of Biology, University of St Andrews, St Andrews KY16 9TH, UK

GDR, 0000-0001-8943-6609

Although circular data are common in biological studies, the analysis of such data is often more rudimentary than it need be. One of the most common hypotheses tested is whether the data suggest that samples are clustered around a certain specified direction, rather than being uniformly spread across all possible directions. Here, I use data from a recent publication on the compass directions of epiphytes and mistletoes on tree trunks. This is used to demonstrate how with relatively little extra work researchers can improve the rigour of testing such hypotheses, and this improved rigour can lead to biological insights missed by simpler analyses. Specifically, I highlight that a much broader range of null hypotheses can be tested than current practice, and that a range of methods are available for estimating a confidence interval for mean direction. I offer advice on appropriate selection for both tests and parameter estimation methods, and highlight the need to correct for the fact that sample estimates are biased estimates of population parameters for circular data.

Circular data occur in many areas of biology, whenever we study anything that is inherently cyclic in nature. The most obvious application is the measurement of anything involving compass bearings or orientations (e.g. the direction of long axes of grazing herbivores with respect to the wind, the compass direction of dispersal of seeds from the parent plant, or the angles of trichomes to the surface of their leaf). The other common application is exploring rhythms in time (e.g. are lions more active as certain times of day, are turtle hatching dates related to specific points in the lunar cycles, is limpet reproduction influenced by the tidal cycle, how can we best describe the distribution of bush fires across the year?). Despite this, very few biologists will have had formal training in handling this type of data, and most statistical textbooks aimed at biologists discuss circular data briefly if at all. Because of this, statistical handling of such data is often more rudimentary than that applied to other types of data. Here, I illustrate this and offer advice on improved practice for a very simple but common issue with circular data: exploring whether a sample of data seems to be clustered around a particular predicted direction. In order to present this in an intuitive form for biologists, I use data from the recent paper of Taylor & Burns [1] exploring the radial distributions of epiphytes and mistletoes around the trunks of trees. I use these data only because the biology of the system is easy to understand, and because the data are very standard. In selecting this paper, I imply no specific criticism of the original work; indeed (while I might have addressed the data slightly differently) I consider that there is nothing fundamentally wrong in their statistical treatment of circular data and their main conclusions would be unaffected by any reasonable treatment of that data.

Taylor & Burns studied the distribution of air plants (species that do not have roots in the soil) around the circumferences on the trunks of beech trees
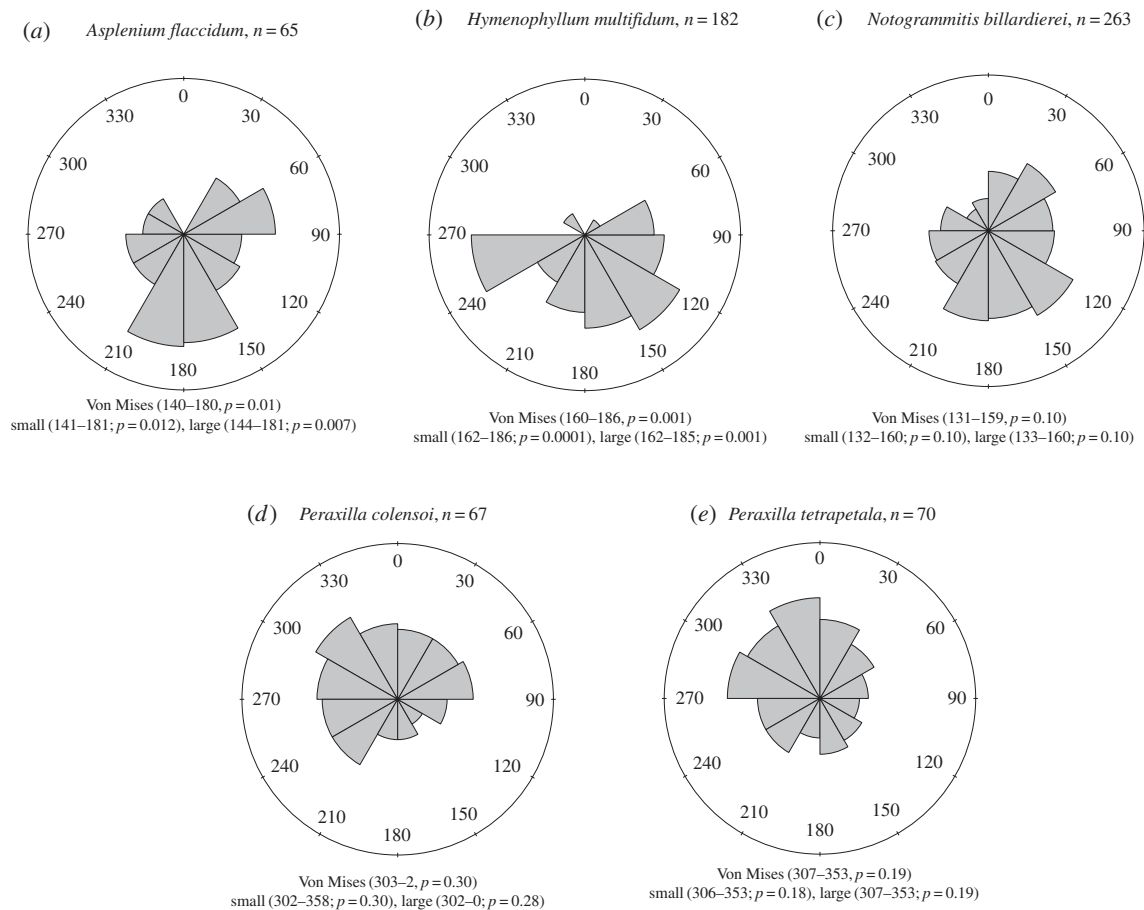
Figure 1. Rose diagrams showing the distributions of orientations of samples of three species of epiphyte and two species of mistletoe. Angles are measured in degrees clockwise from north (0°). Also shown are the sample sizes (*n*) for each and the 95% confidence intervals for the population mean directions generated using three methods: assuming an underlying von Mises distribution, making no assumption about the distribution but using a bootstrapping approach appropriate to small samples, and making no assumption about the underlying distribution but assuming the sample size is sufficient to justify a normal approximation in construction of the confidence interval. For all methods, a *p*-value is generated relating to the null hypothesis that the population mean value has a specified direction: 135° (SE) for epiphytes and 315° (NW) for mistletoes; under the von Mises assumption, a *p*-value was obtained by trial and error as the largest value α such that the $100(1 - \alpha)$ % confidence interval excluded the specified value [2, p. 97].

at their study site in a subalpine forest in the South Island of New Zealand. Two species of mistletoe were common (*Peraxilla colensoi* and *P. tetrapetala*) and three epiphytic ferns (*Asplenium flaccidum*, *Hymenophyllum multifidum* and *Notogrammitis billardierei*). The authors postulated that since the epiphytes use their host trees solely for structural support, their water requirements are met from air humidity and they will predominantly grow on the shady side of trees where humidity is highest. By contrast, mistletoes tap into their hosts for water, and they are predicted to predominantly grow on the sunny side of trees to maximize light gathering for photosynthesis. Taylor & Burns argue that in their study region NW orientated aspects receive most sunlight, and they demonstrate that on trees at heights relevant to their five focal species photosynthetically active radiation is highest on the NW aspect and air humidity highest on the SE aspect. Their measurements for the midpoint of growth of their five focal species on a sample of trees are shown in figure 1.

Given the motivation for the study described above, the logical null hypothesis to test for each of the five focal species is that there is no preferred direction and sample midpoints are uniformly distributed around the trunks of trees. Taylor & Burns test this null hypothesis using the most commonly used test in circular statistics—the simple form of the Rayleigh test; in each case they find evidence to reject the null hypothesis.

In fact, there is a range of alternative tests that could have been used. The simple form of the Rayleigh test makes the assumption that if there is a departure from the null hypothesis that departure has a single peak (it is unimodal). Hence it is appropriate if only unimodal departures are suspected or of interest but the peak of such alternative distributions could occur anywhere around the circle. There is an alternative form of the Rayleigh test that is much less commonly used in the applied literature, where an expected mean direction μ is specified. In this case, the alternate hypothesis of the test is again of a unimodal departure, but now tied to a specific place on the circle, with mean μ. The benefit of adopting this more specific alternative is greater power to reject the null hypothesis where it is false. Both these tests are available within the *circular* package of the R statistical software. For our air-plant example, there is an *a priori* prediction of a unimodal departure from uniformity, and most importantly an expected value of μ (315° corresponding to NW for the two mistletoes and 135° corresponding to SE for the three epiphytes). Hence, I should have recommended using the more specific form of the Rayleigh test in this situation. However, since the less powerful form found significant departures from uniformity for all five species tested, the scientific conclusions from the study would have been unchanged in this case. The two forms of the Rayleigh

test can be shown to be the most efficient test possible [3] if the unimodal alternative is of Von Mises form (the bell-shaped circular analogy of the normal distribution). However it is sufficiently high performing over such a broad range of shapes of unimodal distribution that no alternative test is in common usage [4].

If it cannot be assumed that any departure from uniformity will be unimodal then there are several alternative tests that might be applied (see [2] for an overview of the key literature). Kuiper's *V*, Watsons U^2 and Rao's spacing tests are all available in the *circular* package, and all generally give reliable performance, but none is universally superior to the others. A reasonable approach for biologists would be to apply all three tests. In the great majority of circumstances, the tests will produce very similar *p*-values. If there is notable difference in the *p*-values then visual inspection of the data together with reading of the literature pointed to by Pewsey *et al.* [2] should guide you to the most trustworthy test in your situation.

Returning to the use of a Rayleigh test of the null hypothesis of uniformity, there is an important subtlety in the interpretation of rejection of the null hypothesis. In general, when the null hypothesis is rejected, we want to make some inference as to where the unimodal departure from uniformity is concentrated (i.e. where there is a single peak, we want to identify where that peak is). A common misconception [5] is that if we use the version of the Rayleigh test where μ is specified then rejection of the null hypothesis implies that the underlying population peak is at the specified value of μ.

In their interpretation of the null-hypothesis rejection for the general form of the Rayleigh test where μ is not specified, Taylor & Burns follow the conventional approach and draw conclusions about the population value of μ simply from visual inspection of plots of the data such as figure 1. They conclude that their predictions about mean direction for both mistletoes and epiphytes appear supported by visual inspection. Here, I should like to point out that it is possible to be more quantitative. Firstly, it is relatively straightforward to calculate a confidence interval for the population value of μ, and secondly, it is possible to generate a *p*-value associated with the null hypothesis that the underlying population has a unimodal distribution with the specified value of μ.

Confidence intervals for the population value of μ can be obtained using the *mle.vonmises.bootstrap.ci* function from the R package *circular* if the assumption is made that the underlying distribution is von Mises. The exact method used is detailed in [3]; however, essentially this involves fitting a von Mises function to the data and estimating the two parameters that fully specify the exact shape of the function. One of these parameters is the mean value μ that we are interested in. Bootstrapping is used to give greater reliability than asymptotic results for lower sample sizes, the confidence interval limits being simply the appropriate percentiles of the distribution of bootstrapped estimates of μ. Alternatively, the confidence interval can be calculated without making any assumptions about the nature of the underlying distribution using the method described in [2, pp. 91–96]. This text offers two options: for small samples it recommends confidence interval construction by bootstrapping (with a method derived from one introduced in [6]). Conventional simple bootstrapping is used with the confidence limits being obtained from the appropriate percentiles of the estimates of population means generated from the bootstrap samples.

However, these estimates of population means are not just the simple sample (circular) means but are adjusted to acknowledge that sample estimates of parameters such as the mean direction are known to be biased estimates of the population values [2, p. 90]. For larger samples, the Pewsey *et al.* [2] consider it appropriate to use Z-scores along with bias-corrected estimates of the mean and standard deviation to generate the confidence interval. Pewsey *et al.* [2] also offer methodology (in small- and large-sample forms) for generating a *p*-value associated with the null hypothesis of the population mean direction having a specified value. Both these methods involve calculation of a Z-score test statistic. In the large-sample case, a *p*-value is obtained by comparing this Z-score value to a standard normal distribution; in the small sample case, bootstrap resamples are used, and the *p*-value is the fraction of resamples producing a more extreme Z-score than the original data. As discussed earlier, Taylor & Burns predict that epiphytes should preferentially grow on the aspect of trees offering greatest humidity; in their study region this is SE (135°). Conversely, they predict that mistletoes should concentrate on the strongest lit aspect of trees, which in their region is NW (315°). I applied all these methods to Taylor & Burns' data and the results are provided in figure 1.

A number of observations can be made on these results. Firstly, for all five species the calculated confidence intervals are very similar under all three approaches. This is particularly noteworthy given that none of the five distributions has the classical visual appearance of a von Mises distribution. Like a normal distribution, a von Mises distribution is symmetrical and unimodal, with concentration around the mean falling away to give a bell shape. Next, I note that confidence intervals are relatively wide (between around 35° and 60° depending on species). This again is unsurprising given that none of the species shows a sharp unimodal sample distribution, and broad spread of the underlying distribution makes estimation of its parameters from a sample relatively challenging.

For the two mistletoe species, the calculated *p*-values do not provide evidence to reject the predicted underlying mean direction of 315°, and all confidence intervals include this value, but confidence interval widths are typically around 40° for one species and 60° for the other; thus we have some but not strong support for the predicted NW mean direction (in comparison to NNW for example).

For the three epiphytes, the situation is more complex, for *Notogrammis billardierei* the *p*-values do not give cause to reject the null hypothesis and the confidence intervals include the predicted values. But I again note that the confidence intervals are around 30° in width and the overwhelming majority of values supported within the confidence interval are greater (more southerly) than the predicted value (i.e. SSE rather than SE). For the other two species, the *p*-values do give cause to reject the null hypothesis and the confidence intervals point to a more southerly mean direction than the predicted SE one. Thus, in conclusion there is certainly evidence to suggest that the epiphytes do have different orientation preferences to the mistletoes, and that their preferences is towards the south and east, but generally more towards the south than the east. This conclusion in particular would have been difficult to develop and justify purely from visual inspection of the data.

My purpose here was to demonstrate that most authors do not fully exploit the range of methodologies open to them when analysing circular data, and that when they embrace the full range of techniques available in recent

monographs [2,4,6] and freely available software (https://cran.r-project.org/web/packages/circular/circular.pdf) they can extract further valuable inferences from their data that are currently going undiscovered.

References

1. Taylor A, Burns K. 2016 Radial distributions of air plants: a comparison between epiphytes and mistletoes. *Ecology* **97**, 819–825. (doi:10.1890/15-1322.1)

2. Pewsey A, Neuhäuser M, Ruxton GD. 2013 *Circular statistics in R*. Oxford, UK: Oxford University Press.

3. Watson GS, Williams EJ. 1956 On the construction of significance tests on the circle and the sphere. *Biometrika* **43**, 344–352. (doi:10.1093/biomet/43.3-4.344)

4. Jammalamadaka SR, Sengupta A. 2001 *Topics in circular statistics*, vol. 5. New York, NY: World Scientific.

5. Aneshansley DJ, Larkin TS. 1981 V-test is not a statistical test of 'homeward' direction. *Nature* **293**, 239. (doi:10.1038/293239a0)

6. Fisher NI. 1993 *Statistical analysis of circular data*. Cambridge, UK: Cambridge University Press.