Received: 30 August 2023 | Accepted: 23 February 2024

DOI: 10.1111/2041-210X.14316

Methods in Ecology and Evolution

RESEARCH ARTICLE

# Phenological patterns in ecology: Problems using circular statistics and solutions based on simulations

Michael R. Willig[1,2,3] | Julissa Rojas-Sandoval[3] | Steven J. Presley[1,2,3]

[1]Department of Ecology & Evolutionary Biology, University of Connecticut, Storrs, Connecticut, USA

[2]Center for Environmental Sciences & Engineering, University of Connecticut, Storrs, Connecticut, USA

[3]Institute of the Environment, University of Connecticut, Storrs, Connecticut, USA

**Correspondence**
Michael R. Willig
Email: michael.willig@uconn.edu

**Funding information**
BioXFEL Science and Technology Center, Grant/Award Number: DEB 1950643; National Institutes of Health, Grant/Award Number: AI049725 and TW007120; Division of Environmental Biology, Grant/Award Number: 1831952; University of Connecticut

**Handling Editor:** Phil Bouchet

## Abstract

1. Quantification of phenological patterns (e.g. migration, hibernation or reproduction) should involve statistical assessments of non-uniform temporal patterns. Circular statistics (e.g. Rayleigh test or Hermans-Rasson test) provide useful approaches for doing so based on the number of individuals that exhibit particular activities during a number of time intervals.
2. This study used monthly reproductive activity as an example to illustrate problems in applying circular statistics to data when marginal totals characterize experimental designs (e.g. the number of reproductively active individuals per time interval depends on sampling effort or sampling success). We illustrate the nature of this problem by crafting four exemplar data sets and developing a bootstrapping simulation procedure to overcome complications that arise from the existence of marginal totals. In addition, we apply circular statistics and our bootstrapping simulation to empirical data on the reproductive phenology of six species of Neotropical bats from the Amazon.
3. Because sampling effort or success can differ among time intervals, circular statistics can produce misleading results of two types: those suggesting uniform phenologies when empirical patterns are markedly modal, and those suggesting non-uniform phenologies when empirical patterns are uniform. The bootstrapping simulation overcomes these limitations: the exemplar phenology in which the percentage of reproductively active individuals is modal is appropriately identified as non-uniform based on the bootstrapping approach, and the exemplar phenology in which the percentage of reproductively active individuals is invariant is appropriately identified as uniform based on the bootstrapping approach. The reproductive phenology of each of the six empirical examples is non-uniform based on the bootstrapping approach, and this is true for bats species with unimodal peaks or bimodal peaks.
4. In addition to problems with marginal totals, a review of analyses of phenological patterns in ecology identified two other frequent issues in the application of circular statistics: sampling bias and pseudoreplication. Each of these issues and potential solutions are also discussed. By providing source code for the execution

of the Rayleigh test and Hermans-Rasson test, along with the code for the bootstrapping simulation, we offer a useful tool for assessing non-random phenologies when marginal totals characterize experimental designs.

**KEYWORDS**
binary data, bootstrapping approaches, Hermans-Rasson test, marginal totals, pseudoreplication, Rayleigh test, reproductive phenologies, sampling bias

## 1 | INTRODUCTION

A major thrust of ecological research focuses on characterizing spatial and temporal variation in biologically relevant characteristics and understanding their causes and consequences (Scheiner & Willig, 2011). The need to do so is particularly critical in the Anthropocene, as the rate of temporal change in the environment is accelerating, with increasingly well documented evidence of severe effects on phenological patterns in populations and communities, as well as on the ecosystem services that they mediate in support of human well-being (Beard et al., 2019; Kharouba et al., 2018; Wolkovich et al., 2014). Evaluations of temporal patterns can be undertaken based on two general approaches: (1) those that explore temporal variation by characterizing monotonic, linear, or curvilinear relationships (i.e. much of parametric or non-parametric statistics); or (2) those that explore temporal variation by explicitly incorporating periodicity of temporal attributes (i.e. circular statistics).

The application of approaches based on circular statistics to biological questions in general (e.g. Gustafson & Partch, 2015; Taube, 2007) and to ecological questions in particular (Shimatani et al., 2012; Wiltschko & Wiltschko, 1972) are increasingly common. They have been used to explore a broad diversity of topics: niche-partitioning (e.g. Castro-Arellano et al., 2009, 2010); plant regeneration (e.g. Abe et al., 2012; Aradottir et al., 1997); animal orientation (e.g. Fitak & Johnsen, 2017; Ożarowska et al., 2013); ecophysiology (e.g. Garretson & Forkner, 2021; Pabon-Moreno et al., 2020); and reproduction (e.g. Morellato et al., 2010; Staggemeier et al., 2020).

### 1.1 | Reproductive phenologies

Phenological studies quantify the timing of biological events at any level in the biological hierarchy from individuals to ecosystems (Lieth, 1974). In particular, reproductive phenologies are subject to intense selection pressures as the timing of birth and recruitment of offspring into populations is linked intimately to fitness. Phenologies are constrained by availability of resources or likelihood of multiple interspecific interactions associated with predation, competition, and mutualism. Consequently, considerable research has focused on the description of reproductive phenologies and the development of quantitative approaches for uncovering the nature of annual patterns for plants (e.g. Calero & Rodrigo, 2019; Gomes et al., 2019; Lima et al., 2021; Sheldon & Nadkarni, 2015) and animals (e.g. Eghbali & Sharifi, 2023; Hazard et al., 2022; Lima et al., 2021; Nurul-Ain et al., 2017; Willig, 1985a; Wilson, 1973). Much of the early work on reproductive phenologies was descriptive without formalization of approaches for distinguishing non-modal (uniform) and modal (nonuniform) patterns, or for distinguishing unimodal from multimodal patterns (e.g. Willig, 1985a). More recently, some have incorporated explicit quantitative rules for identifying patterns (e.g. Durant et al., 2013; Willig & Presley, 2023) or have incorporated circular statistical approaches to define phenological patterns that are different from uniform distributions (Hazard et al., 2022; Staggemeier et al., 2020).

### 1.2 | Circular statistics

Circular approaches in statistics for evaluating phenological patterns are required when a focal variable is measured on an interval scale, but the designation of high or low values is arbitrary (e.g. conventions are used to designate magnitude such as noon being represented by "12:00", south being represented by "180" degrees or March being represented by "3"). In those cases, differences in the magnitude of measures do not necessarily correspond to temporal distances (e.g. December [12] is closer to January [1] than it is to June [6]), although differences in monthly values suggest the opposite (i.e. 12–1 = 11 [farther] whereas 12–6 = 6 [closer]). Consequently, descriptive statistics are misleading (e.g. if pregnant females were only detected in December [12] and February [2], the mean monthly value would be 7 {[12 + 2]/2}, indicating July, which is absurd given the concentration of reproductive activity during austral summer months rather than during austral winter months). Data from such circular distributions would not be appropriately executed using classical statistical approaches and require more nuanced representation of the circular scale and distribution of data (Batschelet, 1981; Fisher, 1993; Upton & Fingleton, 1989).

#### 1.2.1 | Data transformation

A multistep process is required to transform phenological data into a form that can be useful for circular statistical analyses. The first step is based on the conversion of times ($Y$) to angular directions ($\alpha$) via

$$\alpha = 360^{\circ}\,(Y / k),$$

where $k$ is the number of units in a full circle (e.g. 24 hours in a day, 12 months in a year, 365 days in a year). The second step represents estimation of magnitude (i.e. the frequency of observed phenological characteristics associated with each time $Y$, such as the number of pregnant females in January represented by angular direction $\alpha = 30°$). In the third step, trigonometric functions (sine and cosine) can be applied to such data to facilitate estimation of the central tendency of angles (means or medians), angular dispersion and confidence intervals. Moreover, such data can be formally incorporated into hypothesis tests regarding the uniformity or modality of angles. Indeed, a growing number of circular statistical tests has been considered for evaluating uniform patterns (null hypothesis) versus non-uniform patterns (alternate hypothesis), with the Rayleigh test (Zar, 2009) and the Hermans-Rasson test (Landler et al., 2018) being used most commonly. Recent assessments of the statistical power of five commonly used circular statistics (i.e. Rayleigh test, V-test, Watson test, Kuiper's test and Rao's Spacing test) led to the general recommendation of using Rayleigh test when the alternative hypothesis is unimodality (i.e. directionality) and an a priori expectation does not exist about mean $\alpha$ (Landler et al., 2018). For multimodal alternative hypotheses, a power assessment led to the conclusion that the Hermans-Rasson test generally, and sometimes substantially, outperforms competing approaches (e.g. Rayleigh test, Rao's Spacing test, Watson test, Kuiper's test). Based on additional comprehensive simulation analyses, Landler et al. (2019) recommended that the Hermans-Rasson test should become the preferred approach because it performs almost as well as the Rayleigh test in unimodal situations, but substantively out-performs the Rayleigh test in multimodal situations. More complex and sometimes powerful analyses (e.g. periodic regression and multivariate analysis of variance) that are based on trigonometric transformations of $\alpha$ via sine or cosine functions amplify the kinds of questions that can effectively be addressed with circular data (Adrian & Meeuwig, 2001; Landler et al., 2022). Regardless of the statistical test, the assumption is that each of $N$ individuals that compose the data set is free to be associated with each particular $\alpha$.

### 1.2.2 | Data nuances

Unfortunately, these approaches in circular statistics are not appropriate when marginal totals (e.g. number of captures per month) arise as a consequence of sampling effort or sampling success, and thereby constrain the number of individuals in any reproductive category for any value of $\alpha$. For example, in some experimental designs involving annual reproductive phenologies (e.g. Durant et al., 2013; Estrada & Coates-Estrada, 2001; Hazard et al., 2022; Nurul-Ain et al., 2017; Willig, 1985a, 1989b; Willig & Presley, 2023) the number of individuals associated with a demographic response variable is actually binary (e.g. pregnant versus not pregnant) and failure to include this inherent binomial characteristic of the response could lead to inaccurate conclusions. In such sampling designs, the number of pregnant females associated with each value of $\alpha$ depends on two data characteristics: (1) the proportion of females in the population that is pregnant during each interval $\alpha$, and (2) the number of individuals, pregnant or not pregnant, that were observed during each interval $\alpha$. Consequently, heterogeneity in sample sizes among intervals, especially if some samples are small for particular values of $\alpha$, is problematic. For example, 0 pregnant females in a month when no females were observed provides no information about reproductive activity during that month, whereas no pregnant females in a month when 100 females were observed is quite informative about reproductive activity during that month. Essentially, circular statistics assume two things. First, a single sample characterizes the data (e.g. the sum of the number of pregnant individuals captured in all months), without monthly marginal subtotals to constrain the number of pregnant females (or non-pregnant females) per month. Second, each individual is free to occur within any time ($\alpha$) during the annual cycle (the analogue of releasing an individual and recording the direction of its movement, where it is "free" to move in any direction). In the data supporting many phenological studies of reproduction, especially for animals (e.g. Durant et al., 2013; Hazard et al., 2022; Willig, 1985a, 1985b; Willig & Presley, 2023), the empirical number of reproductive females associated with a particular monthly interval is constrained by the number of individuals observed during that monthly interval, and is affected by sampling effort and success as well as by the extent to which environmental conditions during a particular monthly interval favour pregnancy.

### 1.2.3 | Objectives

Our goals are to (1) demonstrate via exemplar datasets, how application of classical circular statistics in designs with unequal marginal total can lead to erroneous and counterintuitive conclusions; (2) develop a bootstrap approach to overcome limitations associated with unequal marginal totals; (3) apply the bootstrap approach to exemplar data sets to highlight its salient improvement; and (4) as a proof of concept, apply two circular statistics (i.e. Rayleigh and Hermans-Rasson tests) and the proposed bootstrap approach to reproductive phenologies derived from well-studied Neotropical bats from the Amazon of Peru.

## 2 | MATERIALS AND METHODS

### 2.1 | Exemplars

We designed a suite of exemplar datasets to illustrate the efficacy of circular statistics in detecting phenological patterns. In doing so, we illustrate problems with analysis of phenological uniformity based on numbers of individuals associated with temporal intervals that fail to consider variation in the magnitude of marginal totals for those intervals. More specifically, we constructed four scenarios to reflect the combination of ways that empirical patterns based on percentage of pregnant females (constructed to be modal

or uniform) would be assessed using classical circular statistics. We named such scenarios with a binomial, representing the pattern based on percentages followed by the inferred pattern from statistical circular analyses: Uniform-uniform, modal-modal, uniform-modal and modal-uniform (Figure 1). Separately for the data in each of the four exemplars, we applied two tests based on circular statistics (e.g. Hermans-Rasson test [Landler et al., 2018] or Rayleigh test [Zar, 2009]) to evaluate if phenologies were uniform or not based on the number of pregnant females. Importantly, the alternative hypothesis in the Rayleigh test focuses on directionality (unimodality),

**Uniform-Uniform Pattern**

No Differences Based on N and No Differences Based on %

| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Numbers | Pregnant | 25 | 30 | 25 | 30 | 20 | 25 | 20 | 30 | 20 | 25 | 20 | 30 | 300 |
| | Not Pregnant | 75 | 70 | 75 | 70 | 80 | 75 | 80 | 70 | 80 | 75 | 80 | 70 | 900 |
| | Marginal Total | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1200 |

No Differeces Based on N and No Differences Based on %

| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percents | Pregnant | 25% | 30% | 25% | 30% | 20% | 25% | 20% | 30% | 20% | 25% | 20% | 30% | |
| | Not Pregnant | 75% | 70% | 75% | 70% | 80% | 75% | 80% | 70% | 80% | 75% | 80% | 70% | |
| | Marginal Total | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | |

**Modal-Modal Pattern**

Differeces Based on N and Differences Based on %

| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Numbers | Pregnant | 20 | 20 | 20 | 20 | 20 | 100 | 100 | 20 | 20 | 20 | 20 | 20 | 400 |
| | Not Pregnant | 80 | 80 | 80 | 80 | 80 | 0 | 0 | 80 | 80 | 80 | 80 | 80 | 800 |
| | Marginal Total | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1200 |

Differeces Based on N and Differences Based on %

| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percents | Pregnant | 20% | 20% | 20% | 20% | 20% | 100% | 100% | 20% | 20% | 20% | 20% | 20% | |
| | Not Pregnant | 80% | 80% | 80% | 80% | 80% | 0% | 0% | 80% | 80% | 80% | 80% | 80% | |
| | Marginal Total | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | |

**Uniform-Modal Pattern**

Differeces Based on N but No Differences Based on %

| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Numbers | Pregnant | 40 | 40 | 40 | 40 | 40 | 100 | 100 | 40 | 40 | 40 | 40 | 40 | 600 |
| | Not Pregnant | 40 | 40 | 40 | 40 | 40 | 100 | 100 | 40 | 40 | 40 | 40 | 40 | 600 |
| | Marginal Total | 80 | 80 | 80 | 80 | 80 | 200 | 200 | 80 | 80 | 80 | 80 | 80 | 1200 |

Differeces Based on N but No Differences Based on %

| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percents | Pregnant | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | |
| | Not Pregnant | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | 50% | |
| | Marginal Total | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | |

**Modal-Uniform Pattern**

No Differeces Based on N but Differences Based on %

| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Numbers | Pregnant | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1200 |
| | Not Pregnant | 100 | 100 | 100 | 100 | 100 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 1000 |
| | Marginal Total | 200 | 200 | 200 | 200 | 200 | 100 | 100 | 200 | 200 | 200 | 200 | 200 | 2200 |

No Differeces Based on N but Differences Based on % (Platykurtic)

| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percents | Pregnant | 50% | 50% | 50% | 50% | 50% | 100% | 100% | 50% | 50% | 50% | 50% | 50% | |
| | Not Pregnant | 50% | 50% | 50% | 50% | 50% | 0% | 0% | 50% | 50% | 50% | 50% | 50% | |
| | Marginal Total | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | |

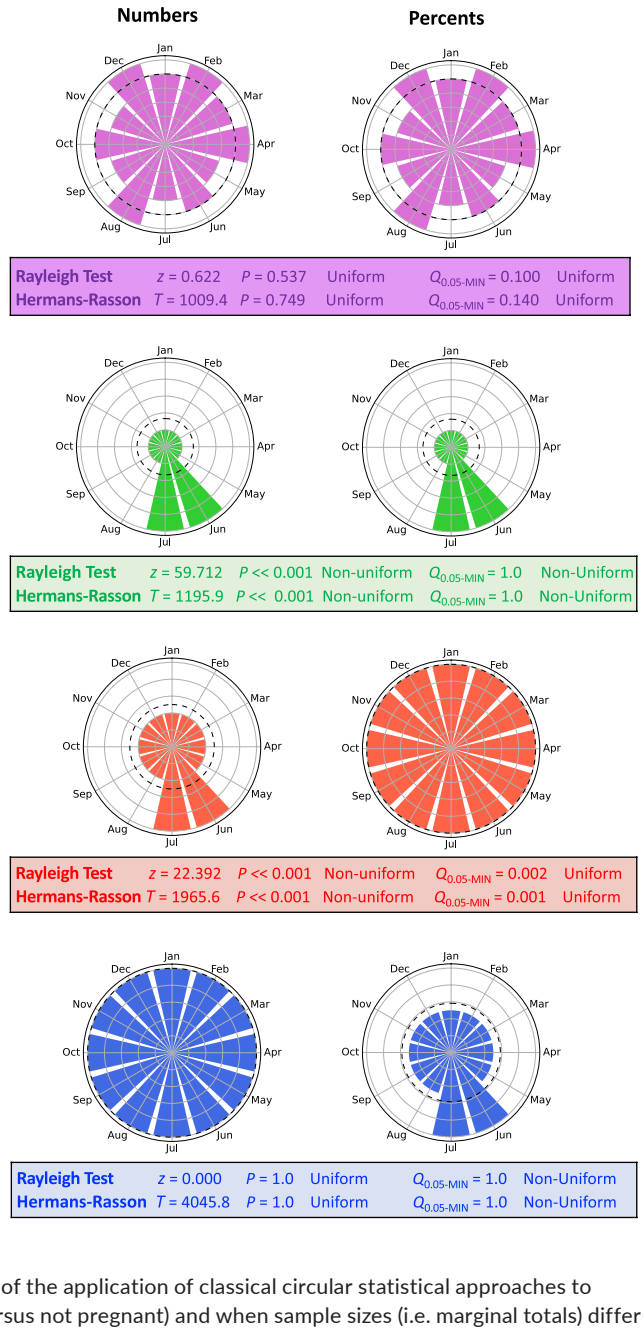



**FIGURE 1** Four exemplar datasets (left column) that illustrate aspects of the application of classical circular statistical approaches to phenologies when focal characteristics are dichotomous (i.e. pregnant versus not pregnant) and when sample sizes (i.e. marginal totals) differ over monthly intervals. The tabular data are expressed in two ways: (1) based on numbers of individuals per monthly interval or (2) based on percentage of individuals that were pregnant per monthly interval. In addition, the tabular data for each exemplar are represented by two wind rose diagrams, one based on numbers and one based on percents. The data based on numbers of individuals per monthly interval were analysed via the Rayleigh test (Zar, 2009) and the Hermans-Rasson test (Landler et al., 2018), as well as by the newly developed bootstrap approach for each of those circular statistics (see text for details). The proportion of simulated *P* values from the bootstrap procedure that were ≤0.05 is indicated by $Q_{0.05}$. For illustrative purposes, we used $N_C = N_{MIN}$ as the basis of results for the more conservative test of uniformity, and $N_C = N_{AVE}$ as the basis of results for a less conservative test of uniformity. Herein, we only report the results for simulations in which $N_C = N_{MIN}$ ($Q_{0.05\text{-}MIN}$), as the results based on $N_C = N_{AVE}$ ($Q_{0.05\text{-}AVE}$) are quite similar and qualitatively identical. The descriptors of the distribution of test statistics and of the *p* values for the Rayleigh test and for the Hermans-Rasson test are reported based on $N_C = N_{MIN}$ as well as based on $N_C = N_{AVE}$ in the Table 1. See the text for details.

whereas the Hermans-Rasson test addresses non-uniformity more broadly, including multi-modality.

The first scenario (Figure 1, uniform-uniform) is characterized by a non-systematically varying number of pregnant females among months (20–30 individuals), with invariant monthly marginal totals (the sum of the number of pregnant and non-pregnant females equal 100). This results in comparable non-systematic variation in the percent of females that were pregnant per month.

The second scenario (Figure 1, modal-modal) is characterized by a variable number of pregnant females over time (i.e. ranging from 20 to 100 individuals), with invariant monthly marginal totals (i.e. 100). This results in considerable variation in the percent of pregnant females per month (20%–100%) with a mode in June–July.

The third scenario (Figure 1, uniform-modal) represents a mismatched pattern that is characterized by a variable number of pregnant females over time (i.e. ranging from 40 to 100 individuals), with the monthly marginal totals being identically variable (ranging from 80 to 200) to maintain the same percent of pregnant females each month. This results in a uniform distribution (i.e. all months are characterized by 50% of the females being pregnant) despite variation among months in observed number of pregnant females due to variation in monthly sample size.

The fourth scenario (Figure 1, modal-uniform) represents a different mismatched pattern in which the number of pregnant females is uniform over time (i.e. 100 individuals), but monthly marginal totals are variable (i.e. 100 to 200). This results in considerable variation in the percent of females that were pregnant per month (50% to 100%) and a peak in percent pregnancy extending for 2 months from June to July. For each of these exemplars, we calculated Rayleigh (*z*) and Hermans-Rasson (*T*) statistics based on a circular statistical test. Rayleigh tests were implemented in the package *Pingouin* (Vallat, 2018). For the Hermans-Rasson test, we developed code following the mathematical definitions provided by Landler et al. (2018, 2019). The *p* values for the Hermans-Rasson tests were obtained by simulation following the mathematical implementation and methodology suggested by Landler et al. (2019). All analyses were performed in Python, and the source codes and used data can be found on Zenodo: https://doi.org/10.5281/zenodo.10799004.

## 2.2 | Simulations

We developed a simple bootstrap approach (Davison & Hinkley, 1997; Efron & Tibshirani, 1994; Wicker, 2021) to overcome the problems of statistical inference when marginal totals act as constraints. We illustrate the approach as a sequence of steps (Figure 2), which we outline hereafter:

1. To remove the effect of variation in marginal totals, establish a common sample size per time interval ($N_C$) that is invariant among intervals. Although no "correct" value for $N_C$ exists, we place reasonable bounds on it by using two values:
   a. $N_{MIN}$, the smallest empirical sample size in the data set (more conservative), or
   b. $N_{AVE}$, the mean sample size in the dataset (less conservative but resulting in a total N that is identical to that in the empirical data).
2. For each time interval *i* with $N_i$ individuals sampled, estimate the number of pregnant females $F_i$ in a bootstrapped population by randomly selecting (with replacement) $N_C$ individuals from the pool of $N_i$ individuals available for that interval.
3. Calculate the test statistic (*z* or *T*) and its significance (*p* value) for the bootstrapped data.
4. Repeat this process 10,000 times to create a distribution of *p* values based on chance when marginal totals are fixed at $N_C$.
5. Explore the distribution of *p* values obtained via the bootstrap approach by estimating:
   a. The mean and median from the distribution of *p* values.
   b. The upper value beyond which 0.025% of simulated *p* values occur.
   c. The lower value beyond which 0.025% of simulated *p* values occur.
   d. The proportion of simulated *p* values ($Q_{0.05}$) that is ≤0.05 (i.e. significant for $\alpha = 0.05$).

We applied this approach to each of the four exemplar data sets (Figure 1). We based decisions about uniformity on the preponderance of information from the distribution of *p* values. More specifically, we consider deviations from uniformity if at least 95% of the simulated *p* values are ≤0.05 (i.e. $Q_{0.05} \geq 0.95$).

## 2.3 | Sensitivity analyses

We conducted preliminary analyses to explore the sensitivity of our simulation approach to: (1) variation in the amplitude of peaks in circular distributions (holding sample size constant) and (2) variation in total sample size (holding proportional peak size constant). We did so for a modified modal-uniform exemplar (Figure 3) based on $N_C = N_{min}$. The effects of peak amplitude on $Q_{0.05}$ were quantified by decreasing peak height from 60 (peak twice the height of background levels) to 30 (peak height indistinguishable from background levels), in intervals of 5% (3 individuals), while maintaining marginal totals for each month. In contrast, the effects of sample size on $Q_{0.05}$ were quantified by proportionally decreasing the marginal totals in intervals by 10%, while maintaining the proportion of pregnant and not pregnant females. For each of these approaches, we conducted evaluations based on the Rayleigh test and the Hermans-Rasson test.

## 2.4 | Applications to reproductive phenologies of bats from Amazonia

Unlike the data that characterized the four exemplar scenarios, variation in reproductive activity and in monthly sample sizes can be considerable in studies of natural populations. This is especially true if data for a number of years of sampling is combined to form

Empirical Data

Iteration Number

| Sampling Month | Reproductive Activity | Monthly Pool | 1 | 2 | . . . | 10,000 |
|---|---|---|---|---|---|---|
| January (1) | Pregant | 50 | 38 | 43 | . . . | 36 |
| | Not Pregnant | 50 | 42 | 37 | | 44 |
| | Total | 100 | 80 | 80 | | 80 |
| February (2) | Pregant | 40 | 44 | 40 | . . . | 39 |
| | Not Pregnant | 40 | 36 | 40 | | 41 |
| | Total | 80 | 80 | 80 | | 80 |
| . . . | | . . . | . . . | . . . | | . . . |
| June (6) | Pregant | 100 | 43 | 45 | . . . | 38 |
| | Not Pregnant | 100 | 37 | 35 | | 42 |
| | Total | 200 | 80 | 80 | | 80 |
| . . . | | . . . | . . . | . . . | | . . . |
| December (12) | Pregant | 45 | 35 | 52 | . . . | 53 |
| | Not Pregnant | 45 | 55 | 38 | | 37 |
| | Total | 90 | 90 | 90 | | 90 |
| Test Statistic (z) | | 22.392 | 14.596 | 12.727 | . . . | 18.010 |
| Significance (P) | | 0.001 | 0.003 | 0.003 | | 0.001 |

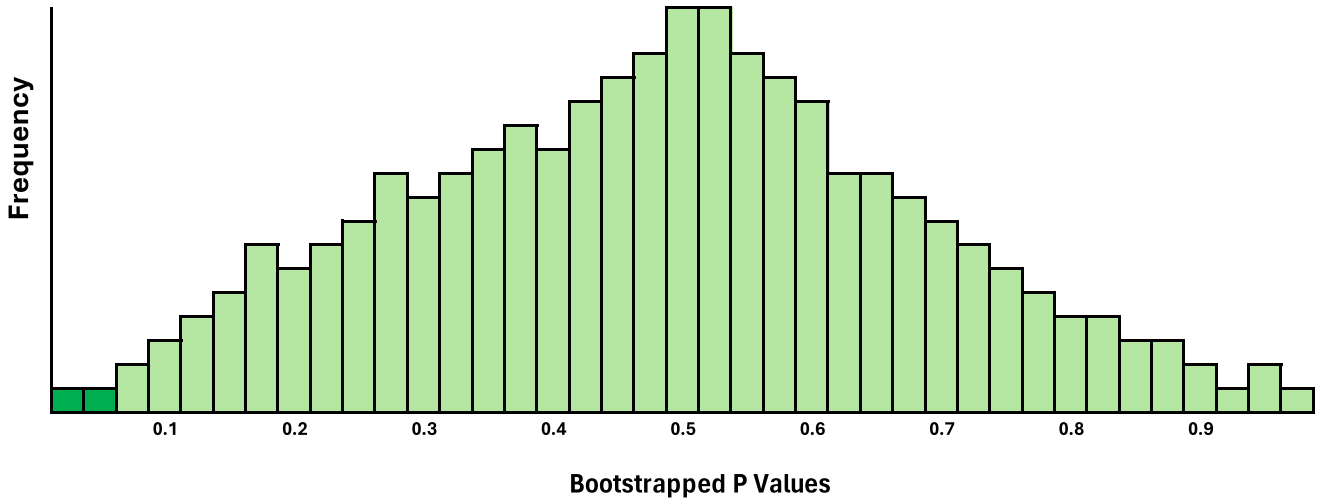


**FIGURE 2** Diagrammatic illustration of the bootstrap simulation approach that was developed to evaluate if the results of classical circular statistical approaches (*p* values) could have arisen by chance and the constraints associated with variation in marginal totals (i.e. sample sizes per monthly interval). Conduct 10,000 iterations of a process whereby the monthly number of reproductively active individuals arises from a random process. To do so (see top of figure), (1) choose a common sample size (i.e. monthly total) for all months in all iterations (e.g. the average monthly total or the minimum monthly total [80, as in this illustration]); (2) use the proportion of reproductively active individuals from each empirical monthly total as the pool (e.g. 0.50 for January) from which a simulated number of reproductively active individuals randomly populates an iteration; (3) calculate a test statistic (e.g. Rayleigh *z* or Hermans-Rasson *T*) and its associated *p* value. Repeat the process 10,000 times to create a distribution of *p* values (see bottom of figure). Consider the phenological pattern to be non-uniform if the percent of simulated *p* values that is ≤0.05 ($Q_{0.05}$) is large (e.g. 0.95). In the illustration, $Q_{0.05} << 0.95$, and the phenological pattern would be considered to be uniform.

the basis for analysis. We used Rayleigh and Hermans-Rasson tests to evaluate monthly variation in reproductive activity for six well-sampled bat populations from the Peruvian Amazon (Willig & Presley, 2023). Without the use of circular statistics, these included reproductive phenologies that have been characterized as (1) unimodal (e.g. *Artibeus planirostris* and *A. obscurus*), (2) bimodal with a brief period of inactivity separating peaks (e.g. *Artibeus lituratus*), (3) bimodal with peaks occurring in tandem (e.g. *Carollia perspicillata* and C. *brevicauda*) and (4) bimodal with diametrically opposed peaks (e.g. *Glossophaga soricina*). Unfortunately, no species exhibited aseasonal reproductive phenologies to serve as an empirical example of uniformity. In addition, we applied the newly developed bootstrap approach to evaluate if the empirical patterns and significance tests based on circular statistics could have arisen as a consequence of variation in marginal totals, and did so when $N_C = N_{MIN}$ as well as when $N_C = N_{AVE}$.

## 3 | RESULTS

### 3.1 | Circular statistics and exemplar data

Conclusions about phenological patterns based on application of circular statistics to data can be misleading because such approaches fail to consider variation in marginal totals, and hence the percent of individuals that are reproductively active per monthly interval (Figure 1).

#### 3.1.1 | Uniform-uniform scenario

As expected based on the construction of the exemplar data, the circular statistical approach did not detect deviations from uniformity for either the Rayleigh test ($z = 0.622$; $p = 0.537$) or the Hermans-Rasson test ($T = 1009.4$; $p = 0.749$). More specifically, the number of pregnant females per month and the percent pregnant females per month varied in a non-systematic manner over time.

#### 3.1.2 | Modal-modal scenario

As expected based on the construction of the exemplar data, the circular statistical approach detected significant deviations from uniformity for both the Rayleigh test ($z = 59.712$; $p << 0.001$) and Hermans-Rasson test ($T = 1195.9$; $p << 0.001$). More specifically, the number and the percent of pregnant females per month were highly modal.

#### 3.1.3 | Uniform-modal scenario

As expected based on the construction of the exemplar data, the circular statistical approach detected significant deviations from uniformity for both the Rayleigh test ($z = 22.392$; $p << 0.001$) and Hermans-Rasson test ($T = 1965.6$; $p << 0.001$), even though the percent pregnant females were constant over time. This erroneous statistical conclusion arose because patterns in the marginal totals created variation in the number of pregnant females per monthly interval.

#### 3.1.4 | Modal-uniform scenario

As expected based on the construction of the exemplar data, the circular statistical approach failed to detect significant deviations from uniformity for both the Rayleigh test ($z = 0.000$; $p = 1.000$) and Hermans-Rasson test ($T = 4045.8$; $p = 1.0$), even though the percent of pregnant females was highly modal. This erroneous statistical conclusion arose because patterns in the marginal totals removed variation in the number of pregnant females among monthly intervals.

### 3.2 | Bootstrap simulations and exemplar data

Application of the bootstrap procedure to the exemplar data correctly identified uniform or modal patterns in phenology in each case, and did so regardless of the decision to use $Q_{0.05\text{-MIN}}$ or $Q_{Q0.05\text{-AVE}}$ (Table 1).

#### 3.2.1 | Uniform-uniform scenario

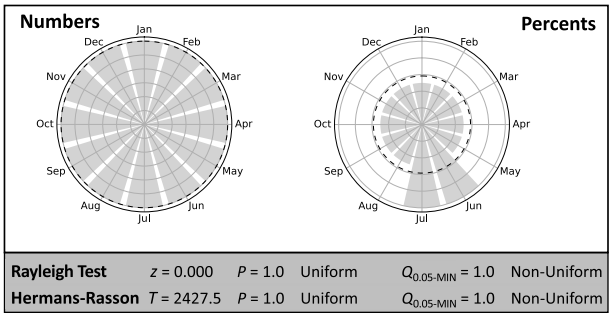
As expected, the bootstrap approach based on the Rayleigh test ($Q_{0.05\text{-MIN}} = 0.100$; $Q_{0.05\text{-AVE}} = 0.099$) and the Hermans-Rasson test($Q_{0.05\text{-MIN}} = 0.140$; $Q_{0.05\text{-AVE}} = 0.143$) corroborated the uniform pattern in reproductive activity (Figure 1). In general, 20%–30% of the females were pregnant during each monthly interval.

(a)

| | No Differeces Based on N but Differences Based on % | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
| Numbers | Pregnant | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 720 |
| | Not Pregnant | 60 | 60 | 60 | 60 | 60 | 0 | 0 | 60 | 60 | 60 | 60 | 60 | 600 |
| | Marginal Total | 120 | 120 | 120 | 120 | 120 | 60 | 60 | 120 | 120 | 120 | 120 | 120 | 1320 |

| | No Differeces Based on N but Differences Based on % (Platykurtic) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
| Percents | Pregnant | 50% | 50% | 50% | 50% | 50% | 100% | 100% | 50% | 50% | 50% | 50% | 50% | |
| | Not Pregnant | 50% | 50% | 50% | 50% | 50% | 0% | 0% | 50% | 50% | 50% | 50% | 50% | |
| | Marginal Total | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | |

(b)




| | | | | | |
|---|---|---|---|---|---|
| Rayleigh Test | $z = 0.000$ | $P = 1.0$ | Uniform | $Q_{0.05\text{-MIN}} = 1.0$ | Non-Uniform |
| Hermans-Rasson | $T = 2427.5$ | $P = 1.0$ | Uniform | $Q_{0.05\text{-MIN}} = 1.0$ | Non-Uniform |

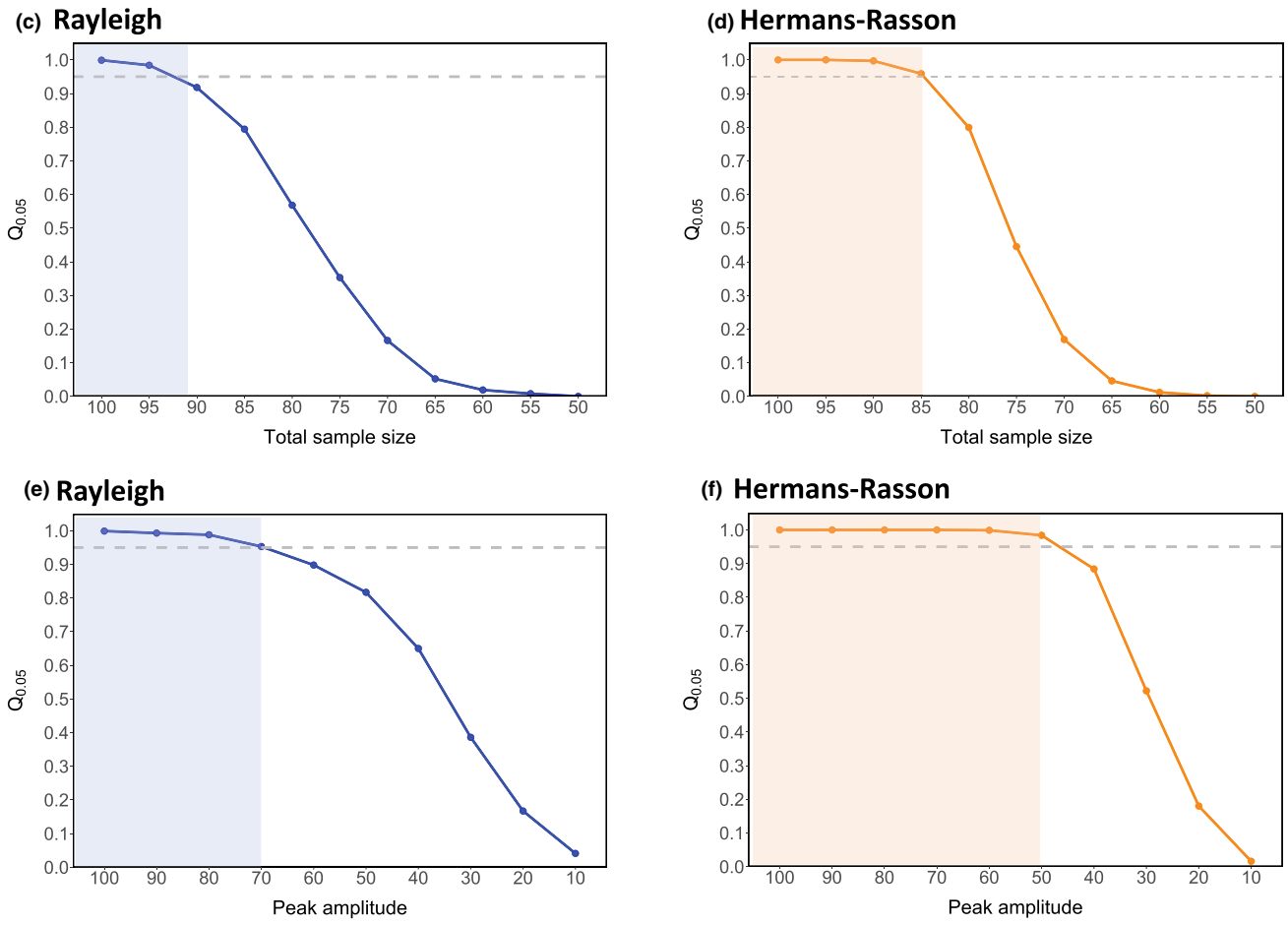



FIGURE 3 Preliminary assessment of the sensitivity of the bootstrap approach to variation in sample size or peak amplitude: (a) Example of phenological data concerning reproductive activity based on numbers or percents used for sensitivity analyses. (b) Graphical representation of the reproductive phenology based on numbers or percents, and the outcome from circular statistics (Rayleigh test and Hermans-Rasson test) that ignore the existence of marginal totals or the inherently binomial nature of the response variable (i.e. pregnant vs. non-pregnant). Results of the bootstrap simulation ($Q_{0.05\text{-min}}$) support the conclusion that phenologies are non-uniform. (c and d) The extent to which $Q_{0.05\text{-min}}$ is sensitive to changes in total sample size for the Rayleigh and Hermans-Rasson tests, respectively. (e and f) The extent to which $Q_{0.05\text{-min}}$ is sensitive to changes in peak amplitude for the Rayleigh and Hermans-Rasson tests, respectively. Horizontal dashed lines indicate thresholds above which non-uniformity characterizes a phenology (i.e. $Q_{0.05\text{-min}} = 0.95$). Shaded areas indicate values of sample size or peak amplitude for which the bootstrap simulation can detect non-uniformity when using either Rayleigh or Hermans-Rasson test.

### 3.2.2 | Modal-modal scenario

As expected, the bootstrap approach based on the Rayleigh test ($Q_{0.05\text{-MIN}} = 1.0$; $Q_{0.05\text{-AVE}} = 1.0$) and the Hermans-Rasson test ($Q_{0.05\text{-MIN}} = 1.0$; $Q_{0.05\text{-AVE}} = 1.0$) corroborated the highly non-uniform pattern in reproductive activity (Figure 1). More specifically, 100% of 100 females were pregnant in June and July, but only 20% of 100 females were pregnant in each of the other 10 months.

### 3.2.3 | Uniform-modal scenario

As expected, the bootstrap approach based on the Rayleigh test ($Q_{0.05\text{-MIN}}=0.002$; $Q_{0.05\text{-AVE}}=0.002$) and the Hermans-Rasson test ($Q_{0.05\text{-MIN}}=0.001$; $Q_{0.05\text{-AVE}}=0.001$) corroborated the uniform pattern in reproductive activity (Figure 1). More specifically, 50% of sampled females in each month were pregnant although sample sizes (200 per month in June and July versus 80 per month in all other intervals) and the number of pregnant females (100 per month in June and July versus 40 per month in all other intervals) differed among months.

### 3.2.4 | Modal-uniform scenario

As expected, the bootstrap approach based on the Rayleigh test ($Q_{0.05\text{-MIN}}=1.0$; $Q_{0.05\text{-AVE}}=1.0$) and the Hermans-Rasson test ($Q_{0.05\text{-MIN}}=1.0$; $Q_{0.05\text{-AVE}}=1.0$) detected deviations from uniformity (Figure 1). That is, two monthly intervals (June and July) represented peaks in the percent of pregnant females (100% of 100 individuals) with all other monthly intervals evincing much lower reproductive activity (50% of 200 individuals).

## 3.3 | Sensitivity analyses

Plots of $Q_{0.05}$ as a function of sample size or as a function of peak amplitude illustrate the sensitivity of the simulation approach to variation in key characteristics associated with empirical phenological studies (Figure 3). Simulations based on the Hermans-Rasson test ($Q_{0.05\text{-HR}}$) were less sensitive than those based on the Rayleigh test ($Q_{0.05\text{-R}}$) when evaluating variation in sample size or in peak amplitude (i.e. mean $Q_{0.05\text{-HR}} \geq Q_{0.05\text{-R}}$). Moreover, metrics were more sensitive to variation in sample size than to variation in peak amplitude (stability in $Q_{0.05}$ [i.e. the plateau in Figure 3f]). For example, peak amplitude can be reduced by >50% (from 100 to less than 50 individuals) before any change in decision about phenology would occur based on the Hermans-Rasson test ($Q_{0.05\text{-HR}} \geq 0.95$). Our sensitivity tests support the conclusions of Landler et al. (2019) in recommending the use of the Hermans-Rasson test because it performs better at detecting non-uniform patterns than does the Rayleigh test when implementing our bootstrapping procedure (Figure 3).

## 3.4 | Reproductive phenologies of bats from Amazonia

In general, conclusions about phenologies of a particular species were the same, whether based on $N_C=N_{MIN}$ or $N_C=N_{AVE}$, or whether based on Rayleigh or Hermans-Rasson test (Table 1).

### 3.4.1 | *Artibeus lituratus*

Based on a quantitative, but non-statistical approach, Willig and Presley (2023) characterized the reproductive phenology of this species to be unimodal (Figure 4). Both the Rayleigh test ($z=7.248$; $p<<0.001$) and Hermans-Rasson test ($T=144.5$; $p<<0.001$) corroborated the modal and non-uniform phenology, respectively, as did the bootstrap approach using statistics based on $N_C=N_{AVE}$ (Rayleigh test, $Q_{0.05\text{-AVE}}=0.999$; Hermans-Rasson test, $Q_{0.05\text{-AVE}}=1.0$). In contrast, the simulations based on $N_C=N_{min}$ (Rayleigh test, $Q_{0.05\text{-MIN}}=0.740$; and Hermans-Rasson test, $Q_{0.05\text{-MIN}}=0.854$) were not sufficiently powerful to detect a non-uniform pattern, likely because the minimum marginal total (April) was quite small (8).

### 3.4.2 | *Artibeus obscurus*

Based on a quantitative, but non-statistical approach, Willig and Presley (2023) characterized the reproductive phenology of this species to be unimodal (Figure 4). Both the Rayleigh test ($z=17.483$; $p<<0.001$) and Hermans-Rasson test ($T=237.6$; $p<<0.001$) corroborated the modal and non-uniform phenology, respectively, as did the bootstrap approach based on either statistic (Rayleigh test, $Q_{0.05\text{-MIN}}=0.999$, $Q_{0.05\text{-AVE}}=1.0$; and Hermans-Rasson test, $Q_{0.05\text{-MIN}}=0.997$, $Q_{0.05\text{-AVE}}=1.0$).

### 3.4.3 | *Artibeus planirostris*

Based on a quantitative, but non-statistical approach, Willig and Presley (2023) characterized the reproductive phenology of this species to be unimodal (Figure 4). Both the Rayleigh test ($z=44.620$; $p<<0.001$) and Hermans-Rasson test ($T=497.9$; $p<<0.001$) corroborated the modal and non-uniform phenology, respectively, as did the bootstrap approach based on either statistic (Rayleigh test, $Q_{0.05\text{-MIN}}=1.0$, $Q_{0.05\text{-AVE}}=1.0$; and Hermans-Rasson test, $Q_{0.05\text{-MIN}}=1.0$, $Q_{0.05\text{-AVE}}=1.0$).

### 3.4.4 | *Carollia brevicauda*

Based on a quantitative, but non-statistical approach, Willig and Presley (2023) characterized the reproductive phenology of this species to be bimodal with tandem peaks (Figure 4). Both the Rayleigh test ($z=44.487$; $p<<0.001$) and Hermans-Rasson test ($T=669.5$; $p<<0.001$) corroborated the modal and non-uniform phenology, respectively, as did the bootstrap approach based on either statistic (Rayleigh test, $Q_{0.05\text{-MIN}}=1.0$, $Q_{0.05\text{-AVE}}=1.0$; and Hermans-Rasson test, $Q_{0.05\text{-MIN}}=1.0$, $Q_{0.05\text{-AVE}}=1.0$).

**TABLE 1** Descriptive characteristics of the distribution of results (test statistics: *z* for the Rayleigh test; or *T* for the Hermans-Rasson test) from the bootstrap simulations for each of the four exemplar datasets and for each of the six empirical datasets (see the text for details). The simulations were conducted for each combination of statistical test (Rayleigh versus Hermans-Rasson) and standardized marginal total ($N_{AVE}$ vs. $N_{MIN}$).

| | | Rayleigh test | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $N_C = N_{AVE}$ | | | | | $N_C = N_{MIN}$ | | | | |
| | | Mean | Median | 2.5% | 97.5% | $Q_{0.05\text{-AVE}}$ | Mean | Median | 2.5% | 97.5% | $Q_{0.05\text{-MIN}}$ |
| Exemplars | | | | | | | | | | | |
| Uniform-uniform | *z* | 1.350 | 1.031 | 0.041 | 4.517 | | 1.349 | 1.010 | 0.040 | 4.533 | |
| | *p* | 0.406 | 0.357 | 0.011 | 0.960 | 0.099 | 0.407 | 0.365 | 0.011 | 0.961 | 0.100 |
| Modal-modal | *z* | 60.26 | 60.07 | 46.76 | 74.51 | | 60.18 | 59.96 | 46.92 | 74.40 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| Uniform-modal | *z* | 0.491 | 0.336 | 0.012 | 1.799 | | 0.499 | 0.344 | 0.012 | 1.855 | |
| | *p* | 0.671 | 0.715 | 0.165 | 0.988 | 0.002 | 0.668 | 0.709 | 0.157 | 0.988 | 0.002 |
| Modal-uniform | *z* | 24.73 | 24.55 | 17.49 | 32.87 | | 13.68 | 13.51 | 8.497 | 19.84 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| Bats | | | | | | | | | | | |
| *A. lituratus* | *z* | 9.772 | 9.656 | 5.190 | 15.01 | | 4.034 | 3.939 | 1.295 | 7.360 | |
| | *p* | <0.001 | <0.001 | <0.001 | 0.005 | 0.999 | 0.047 | 0.017 | <0.001 | 0.278 | 0.740 |
| *A. obscurus* | *z* | 20.34 | 20.27 | 12.73 | 28.45 | | 11.07 | 10.93 | 5.656 | 17.20 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | 0.003 | 0.999 |
| *A. planirostris* | *z* | 58.94 | 58.85 | 46.77 | 71.58 | | 27.55 | 27.38 | 19.29 | 36.49 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| *C. brevicauda* | *z* | 42.35 | 42.22 | 31.07 | 54.46 | | 16.00 | 15.80 | 9.325 | 23.70 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| *C. perspicillata* | *z* | 155.3 | 155.1 | 133.9 | 177.3 | | 90.47 | 90.47 | 74.01 | 107.3 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| *G. soricina* | *z* | 3.607 | 3.331 | 1.010 | 7.664 | | 1.750 | 1.507 | 0.218 | 4.681 | |
| | *p* | 0.073 | 0.035 | <0.001 | 0.366 | 0.581 | 0.276 | 0.223 | 0.008 | 0.807 | 0.136 |
| | | **Hermans-Rasson test** | | | | | | | | | |
| | | $N_C = N_{AVE}$ | | | | | $N_C = N_{MIN}$ | | | | |
| | | Mean | Median | 2.5% | 97.5% | $Q_{0.05\text{-AVE}}$ | Mean | Median | 2.5% | 97.5% | $Q_{0.05\text{-MIN}}$ |
| Exemplars | | | | | | | | | | | |
| Uniform-uniform | *T* | 1007.3 | 1006.3 | 910.4 | 1106.6 | | 1007.2 | 1007.1 | 907.8 | 1107.2 | |
| | *p* | 0.335 | 0.280 | <0.001 | 0.890 | 0.143 | 0.334 | 0.280 | <0.001 | 0.890 | 0.140 |
| Modal-modal | *T* | 1194.1 | 1194.6 | 1091.0 | 1296.5 | | 1194.4 | 1194.8 | 1091.9 | 1297.5 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| Uniform-modal | *T* | 2021.4 | 2021.4 | 1907.1 | 2136.6 | | 1616.9 | 1617.1 | 1515.3 | 1718.2 | |
| | *p* | 0.807 | 0.880 | 0.290 | 1.0 | 0.001 | 0.806 | 0.880 | 0.290 | 1.0 | 0.001 |
| Modal-uniform | *T* | 4256.4 | 4256.4 | 4105.9 | 4407.2 | | 2325.0 | 2324.6 | 2216.1 | 2434.5 | |
| | *P* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| Bats | | | | | | | | | | | |
| *A. lituratus* | *T* | 145.5 | 145.6 | 112.8 | 179.2 | | 54.27 | 54.11 | 33.88 | 75.79 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | 0.026 | 0.011 | <0.001 | 0.151 | 0.854 |
| *A. obscurus* | *T* | 233.0 | 233.1 | 188.5 | 278.7 | | 123.1 | 122.9 | 90.26 | 157.0 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | 0.001 | <0.001 | <0.001 | 0.012 | 0.997 |

TABLE 1 (Continued)

| | | Hermans-Rasson test | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $N_C = N_{AVE}$ | | | | | $N_C = N_{MIN}$ | | | | |
| | | Mean | Median | 2.5% | 97.5% | $Q_{0.05\text{-}AVE}$ | Mean | Median | 2.5% | 97.5% | $Q_{0.05\text{-}MIN}$ |
| *A. planirostris* | *T* | 529.4 | 529.6 | 462.9 | 595.9 | | 244.2 | 243.9 | 199.0 | 289.9 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| *C. brevicauda* | *T* | 649.3 | 648.9 | 575.0 | 725.5 | | 239.7 | 239.4 | 194.4 | 286.2 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| *C. perspicillata* | *T* | 1702.0 | 1701.0 | 1579.4 | 1827.3 | | 987.9 | 987.7 | 892.0 | 1084.0 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 |
| *G. soricina* | *T* | 178.7 | 178.4 | 143.3 | 216.2 | | 68.90 | 68.51 | 46.69 | 92.45 | |
| | *p* | <0.001 | <0.001 | <0.001 | <0.001 | 1.0 | 0.002 | <0.001 | <0.001 | 0.019 | 0.995 |

Abbreviations: 2.5% represents the value below which 2.5% of the simulated values occurs; 97.5% represents the value above which 2.5% of the simulated values occur; $Q_{0.05\text{-}AVE}$ estimates the proportion of simulated *p* values that were ≤0.05 when $N_C = N_{AVE}$; and $Q_{0.05\text{-}MIN}$ estimates the proportion of simulated *p* values that were ≤0.05 when $N_C = N_{MIN}$.

### 3.4.5 | *Carollia perspicillata*

Based on a quantitative, but non-statistical approach, Willig and Presley (2023) characterized the reproductive phenology of this species to be bimodal with tandem peaks (Figure 4). Both the Rayleigh test ($z = 145.479$; $p << 0.001$) and Hermans-Rasson test ($T = 1618.5$; $p << 0.001$) corroborated the modal and non-uniform phenology, respectively, as did the bootstrap approach based on either statistic (Rayleigh test, $Q_{0.05\text{-}MIN} = 1.0$, $Q_{0.05\text{-}AVE} = 1.0$; and Hermans-Rasson test, $Q_{0.05\text{-}MIN} = 1.0$, $Q_{0.05\text{-}AVE} = 1.0$).

### 3.4.6 | *Glossophaga soricina*

Based on a quantitative, but non-statistical approach, Willig and Presley (2023) characterized the reproductive phenology of this species to be bimodal with peaks in September and January–February (Figure 4). The Rayleigh test ($z = 0.378$; $p = 0.687$) failed to detect deviations from uniformity in favour of directionality. In contrast, the Hermans-Rasson test ($T = 145.7$; $p << 0.001$) corroborated the non-uniform phenology. Similarly, the bootstrap approach based on the Rayleigh test ($Q_{0.05\text{-}MIN} = 0.136$; $Q_{0.05\text{-}AVE} = 0.581$) indicated a uniform phenology, whereas the bootstrap approach based on the Hermans-Rasson test ($Q_{0.05\text{-}MIN} = 0.995$; $Q_{0.05\text{-}AVE} = 1.0$) detected a non-uniform phenology.

## 4 | DISCUSSION

Although analyses of reproductive phenology in plants and animals are not uncommon, they are typically derived from studies in which the primary goal of the original research was not related to reproductive biology (e.g. studies of community ecology, biodiversity, conservation or systematics). Moreover, the annual cycle of events in those circumstances is sometimes reconstructed from data that span many annual cycles or that do not represent random samples. Consequently, sampling designs in many cases are not optimal for analysis via circular statistics for a variety of reasons related to marginal totals, pseudoreplication or sampling bias.

We clearly illustrate that the application of circular statistics in designs characterized by marginal totals can lead to counterintuitive and inaccurate results in which uniform patterns based on percentages are identified as non-uniform and modal patterns based on percentages are identified as uniform (Table 1; Figure 1). These mismatches arise because sampling effort or sampling success in experimental designs can affect the number of reproductively active individuals observed during a time interval as much as can the monthly proportion of reproductively active individuals in the population. In addition, we show that a bootstrap simulation can overcome these issues and can appropriately identify both uniform and non-uniform patterns with considerable certainty based on either of two statistics (Rayleigh *z* or Hermans-Rasson *T*) and either of two sample size options applied in the bootstrap, the minimum ($N_{MIN}$) or the average ($N_{AVE}$) monthly subtotal (Table 1; Figure 1). Our preliminary analysis illustrates the effect of changes in sample size or in peak amplitude on the sensitivity of the simulation approach, suggesting that conclusions based on simulations involving the Hermans-Rasson test may be more robust, and that analyses are more sensitive to variation in sample size than to variation in peak amplitude. Nonetheless, such conclusions would depend on the nature of variation during periods of reduced activity.

### 4.1 | Establishment of a common sample size for all time intervals

Establishing a common sample size for all time intervals ($N_C$) in the context of a bootstrapping simulation can overcome problems associated with marginal totals, but the magnitude of the common sample size is necessarily arbitrary. Nonetheless,

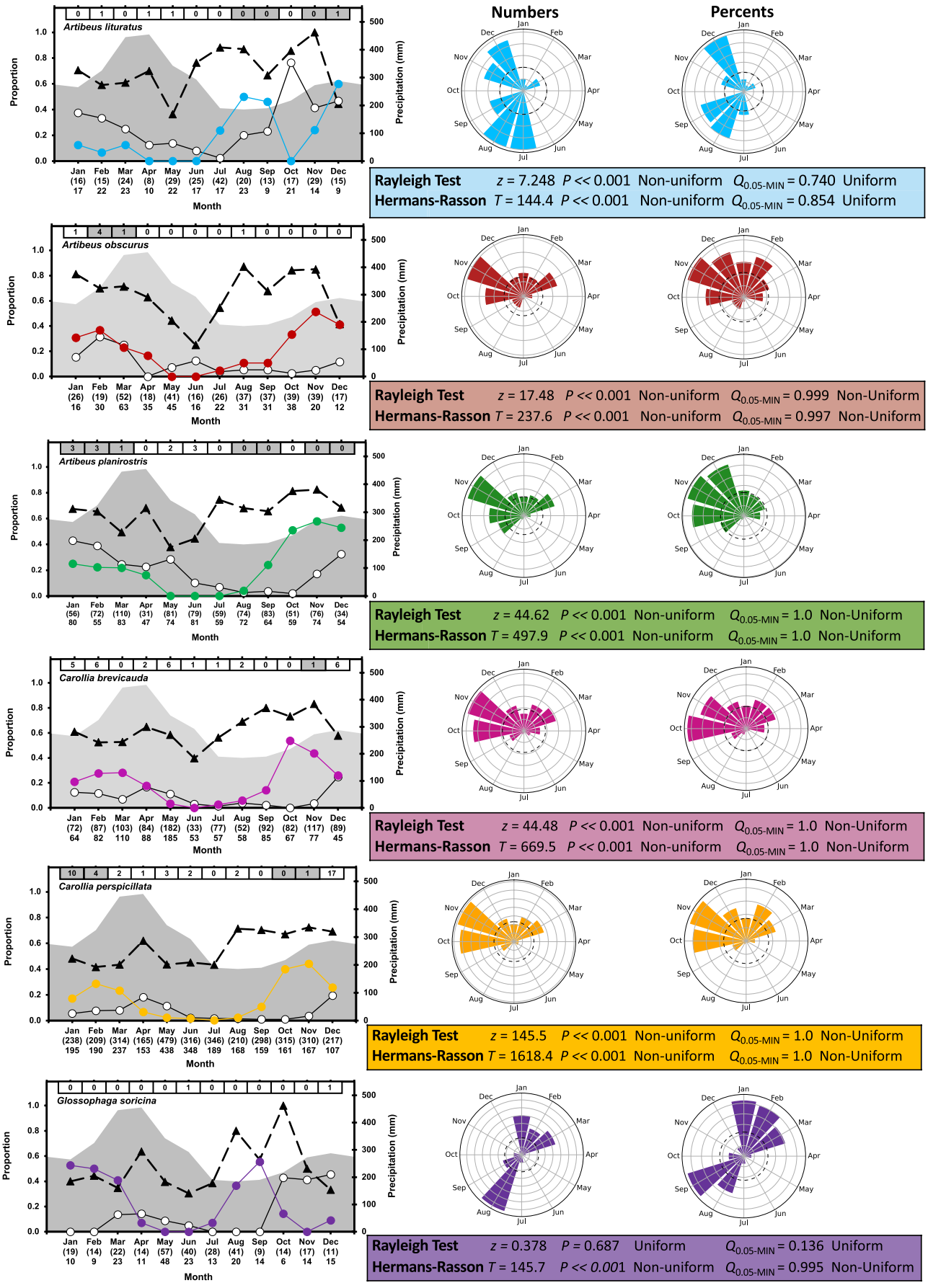


**FIGURE 4** Graphical representations (left column) of reproductive phenologies of each of six species of bats (*Artibeus lituratus*, *Artibeus obscurus*, *Artibeus planirostris*, *Carollia brevicauda*, *Carollia perspicillata* and *Glossophaga soricina*) from the Peruvian Amazon (Willig & Presley, 2023). Shaded background represents mean annual pattern of precipitation for the study area (Iquitos, Perú) for the quinquennium during which bats were captured. Coloured circles represent monthly proportions of adult females that were pregnant. White circles represent monthly proportions of adult females that were lactating. Total numbers of adult females (bases for previous proportions) are indicated by parenthetical numbers below each month. Black triangles represent monthly proportions of adult males with scrotal testes. Total numbers of adult males (bases for previous proportions) are indicated by numbers below each month that are not in parentheses. Numbers in the horizontal row above each graph represent the number of captured juveniles; shaded boxes in that row represent months during which some adult females were simultaneously pregnant and lactating. In addition, the data for each species are represented by two wind rose diagrams, one based on numbers (left) and one based on percents (right). The data based on numbers of individuals per monthly interval were analysed via the Rayleigh test (Zar, 2009) and the Hermans-Rasson test (Landler et al., 2018), as well as by the newly developed bootstrap approach for each of those circular statistics (see text for details). The proportion of simulated *p* values from the bootstrap procedure that were ≤0.05 is indicated by $Q_{0.05}$. For illustrative purposes, we used $N_C = N_{MIN}$ as the basis of results for the more conservative test of uniformity, and $N_C = N_{AVE}$ as the basis of results for a less conservative test of uniformity. Herein, we only report the results for simulations in which $N_C = N_{MIN}$ ($Q_{0.05\text{-}MIN}$) as the results based on $N_C = N_{AVE}$ ($Q_{0.05\text{-}AVE}$) are quite similar. The descriptors of the distribution of test statistics and of the *p* values for the Rayleigh test and for the Hermans-Rasson test are reported based on $N_C = N_{MIN}$ as well as based on $N_C = N_{AVE}$ in the Table 1. See the text for details.

we recommended the use of two approaches, $N_C = N_{MIN}$ and $N_C = N_{AVE}$, to put reasonable bounds on the detection of non-uniformity that are associated with empirical aspects of the data. The use of $N_{MIN}$ is conservative as it rarefies all samples to the minimum empirical sample size, resulting in lower analytical power and, consequently, high confidence in any detections of deviations from uniformity. In contrast, the $N_{AVE}$ is a less conservative value that sets the total sample size of the simulations to perform in a way that is sensitive to all interval subtotals. When both approaches give rise to the same conclusion, suggesting non-uniformity ($Q_{0.05\text{-}MIN} \geq 0.95$ and $Q_{0.05\text{-}AVE} \geq 0.95$) or suggesting uniformity ($Q_{0.05\text{-}MIN} \leq 0.95$ and $Q_{0.05\text{-}AVE} < 0.95$), considerable confidence characterizes the identification of phenological patterns. When the two metrics are not in accord, then results are equivocal, and more careful exploration of the full distribution of *p* values that derive from the simulation (e.g. Table 1) may be necessary or desirable. Although this approach may give rise to uncertainty and appear to be less rigorous than more conventional statistical approaches, it is no more uncertain or arbitrary than using values of $\alpha$ equal to 0.05 versus 0.10 or 0.01 in analyses. Indeed, blind reliance on $\alpha$ as the indicator of significance can be problematic (e.g. is 0.05001 meaningfully less significant than 0.05000), as has been discussed by many others (e.g. De Valpine, 2023; Lakens, 2015; Muff et al., 2022; and in a Special Feature in Ecology [see Ellison et al., 2014]).

## 4.2 | Independence and hidden marginal totals

Some research using circular statistics violates the assumption of independence among observations, even when variation among time intervals in sampling effort or sampling success is not problematic. For example, phenological studies of plants are sometimes based on designs in which a fixed number of individuals (*F*) along a transect or within a plot, is sampled once during each time interval, and the number of individuals in each interval that are "reproductively active" (e.g. flowering or fruiting) represents the data analysed by circular statistics (Lima et al., 2012; Valentin-Silva et al., 2018). This procedure is biased and includes a form of pseudoreplication, as the same *F* individuals are monitored each time interval. This procedure enhances the likelihood that adjacent intervals will have similar numbers of active individuals (e.g. if an individual fruits in May, it may be more likely to fruit in June), thereby potentially overrepresenting directionality in the response over time. This problem is exacerbated when the duration of a particular activity is large compared to the period spanned by each time interval. In these circumstances, a hidden marginal total essentially characterizes each time interval (i.e. the sum of the active and inactive individuals per time interval) that is ignored in analyses. This approach is further complicated by the repeated measurement of activity for each of the *F* individuals (i.e. the number of reproductively active individuals that are registered during all time intervals exaggerates the number of independent observations, as the same individuals can be reproductively active in multiple time intervals). In this circumstance, the total sample size is inflated as is the power of the test.

An alternative metric avoids the problems of pseudoreplication when dealing with designs characterized by such hidden marginal totals. Rather than using the number of reproductively active individuals per time interval to inform circular statistics, using the time of earliest activity by each of the *F* individuals (or the median time period during which each individual was reproductively active) should be used. In this scenario, each of *F* individuals only appears in one time interval, thereby eliminating problems associated with pseudoreplication and marginal totals and obviating the need to apply bootstrap simulations to distinguish uniform and non-uniform patterns. Nonetheless, an additional potential hidden marginal total remains in such scenarios if some individuals do not reproduce during the entire study period. Although we use examples of repeated transect surveys of plants (e.g. Rother et al., 2022) to demonstrate the problem of pseudoreplication, this problem also manifests if movements of the same individuals are used as independent events in radio-tracking studies (e.g. Lorch et al., 2005; Ossi et al., 2020) or if the calls or songs from the same

individuals are used as if they are independent during acoustic monitoring surveys (e.g. Boullhesen et al., 2023).

### 4.3 | Bat reproduction in lowland Amazonia

The application of bootstrap simulations to circular analyses of datasets of bat species shows that this approach successfully overcomes limitations related to marginal totals and that it could be a useful tool for assessing phenological patterns when marginal totals characterize experimental designs. These results provide statistical rigour and support for most of the conclusions of Willig and Presley (2023) regarding the reproductive phenology of bats based on a quantitative approach (Table 1, Figure 4). For 4 of the 6 species (*A. obscurus*, *A. planirostris*, *C. brevicauda* and *C. perspicillata*), non-uniformity characterized the reproductive pattern regardless of metric or $N_C$, reaffirming the conclusions based on quantitative rules of thumb. The reproductive phenology of *G. soricina* detected by the bootstrapping approach depended on the choice of metric but not on considerations of $N_C$. The inability of the bootstrapping approach based on the Rayleigh test to detect non-uniform patterns in which modes are offset by about 6 months (diametrically opposed bimodal circular data) reflects the nature of the Rayleigh test and its alternative hypothesis (directionality) rather than issues with the bootstrapping approach. Fortunately, the Hermans-Rasson test provides a more general assessment of non-uniformity, as the alternate hypothesis does not specify directionality, although it does capture non-uniform results that are directional (*A. obscurus*, *A. planirostris*, *C. brevicauda* and *C. perspicillata*). Finally, the results for the bootstrapping simulation for *A. lituratus* (Figure 4) illustrate the lack of power when $N_C$ is small (i.e. $N_C = 8$ based on the minimum monthly marginal total) regardless of test statistic and the greater power when $N_C$ is large (i.e. $N_C = 21$ based on the average monthly marginal total). Taken together, these results suggest that the data cannot distinguish between uniform and non-uniform patterns with sufficient confidence for *A. lituratus*, although we suspect that the $Q_{0.05\text{-MIN}}$ is quite conservative as seven of the twelve monthly marginal totals were ≤20 individuals (i.e. 8, 13, 15, 15, 16, 17, 20) and the smallest monthly marginal total (April) was quite small (8). Indeed, if $N_C = 10$ for this analysis, then $Q_{0.05\text{-}10} > 0.95$, indicating non-uniform reproductive phenology for *A. lituratus*.

### 4.4 | Other examples with problematic designs

Different studies that apply circular statistics lead to unreliable results because of problems associated with unequal sampling effort among intervals, marginal totals that are apparent or hidden, or unintended biases in sampling. Analytical problems associated with some of these issues can be addressed via our bootstrapping simulation approach (i.e. those involving marginal totals), or by modifying dependent variables to avoid pseudoreplication. In contrast, solutions are unapparent for other kinds of sampling designs, especially those characterized by sampling biases.

Many types of data are inherently binary and suffer from shortcomings that are similar to those associated with the assessment of reproductive phenology. Studies of the effects of the lunar cycle on daily activity patterns illustrates the problem. Lovari et al. (2017) studied the suburban ecology of porcupines, and evaluated whether individuals avoided being active during intervals with considerable illumination from the moon (lunar phobia). To do so, the behaviour (active versus inactive) of each of a number of individuals was determined during each of many instances when those individuals were detected via radio telemetry (i.e. a "fix"). Importantly, each individual was represented by multiple fixes, and the number of fixes was variable among individuals. More specifically, they analysed whether number of "fixes" when porcupines were inactive was homogenous throughout the lunar cycle (i.e. among four temporal categories defined in reference to the phase of the moon) based on Rayleigh's test, and separately did so for each of four seasons. This approach is problematic because of hidden marginal totals (sum of the number of "fixes" during which individuals were active or inactive) for each of the four lunar phases. Moreover, the data suffer from pseudoreplication, as porcupines (each of 11 tracked individuals) were represented by many fixes during each season (i.e. 492 fixes in autumn, 478 fixes in winter, 483 fixes in spring and 512 fixes in summer) and during each time interval. Although, a bootstrap approach could rectify issues with hidden marginal totals, issues concerning pseudoreplication cannot be remedied so easily.

In a transect-based study of fruiting phenology, individual plants (trees) or observation areas (herbs) were monitored on a monthly basis over a 2-year period (Cortés-Flores et al., 2013). For each species, months in which substantial fruiting occurred defined its phenophase. For a monthly interval to be part of the phenophase of a particular species, the monthly interval must be characterized by at least 10 fruiting individuals (woody plants) or at least 10 observation areas that harboured plants with mature fruits (herbaceous plants). Fruiting phenology was based on the number of species per month (a community-level metric rather than a population-level metric) that met or exceeded the threshold for phenophase for each of six groups defined by a combination of dispersal syndrome (anemochorous, autochorous or zoochorous species) and year (November 2007–October 2008 versus November 2008–October 2009). These data are characterized by a lack of independence (and inflated power), as the same individuals of each species were observed throughout all the sampling intervals, with more abundant species able to meet the phenophase threshold more easily than could more rare species. For example, 100% of the individuals in a rare species ($N = 10$) must be fruiting in a particular month for it to meet the phenophase threshold, whereas 10% of the individuals in an abundant species ($N = 100$) need only be fruiting in a particular month for it to meet the phenophase threshold. As a consequence, considerable interspecific differences characterize the potential to contribute to phenological patterns of any dispersal syndrome in any year. In this situation, an alternative solution exists for addressing fruiting phenologies that essentially weights each species by the number of individuals present in the study area. This alternative would base fruiting phenology on individual-level activity (ignoring species identities within

syndromes), using the first or median month in which each individual fruits as the data for analysis by circular statistics. This solution does not suffer from pseudoreplication or hidden marginal totals, but still reflects unequal contributions among species. However, unlike the phenophase approach, the exact contribution of each species to a group's phenology is known and equal to the proportional abundance of the species in the group.

Unfortunately, designs exist for which no known solution can facilitate the appropriate use of circular statistics to evaluate temporal or spatial patterns. For example, herbarium specimens were used to document long-term phenological changes in flowering due to climate change via circular statistics (Lima et al., 2021). Unfortunately, herbarium specimens do not represent random samples of the phenological state of plants during any time interval. Rather, herbarium specimens are likely biased to over-represent individuals that are fruiting or flowering, as these conditions maximize the museological value of the specimen (Alexiades, 1996). At all-time intervals, this effectively results in an over-estimation of the number of reproductively active individuals and an underestimation of the number of non-reproductively active individuals. Consequently, the proportion of reproductively active individuals at any time interval is biased, and the actual marginal totals are unknown. Interspecific comparisons of phenology are further constrained by nuances of experimental design associated with species-specific biases in estimates of reproductive activity and hidden marginal totals. Compared to their relative abundances, uncommon species may be over-represented in museum collections, whereas dominant species may be under-represented (i.e. sampling success differs among species based on their abundance). All other things being equal, less abundant species are more likely to be collected when they are encountered in the field, regardless of reproductive activity (less biased data for assessing phenology), whereas reproductively active individuals are more likely the targets for collection for abundant species. Such species-specific biases would compromise the conclusions about differences among species in reproductive phenologies.

Data collected by citizen scientists is an increasingly important resource for characterizing long-term temporal changes in populations and communities (Cooper et al., 2014). However, such data (e.g. those collected by iNaturalist [https://www.inaturalist.org/]) do not represent standardized effort among time intervals or unbiased observations of targeted biological activities such as reproductive status (Bird et al., 2014). For example, a study on diurnal activity and reproductive phenology of anurans in Brazil based on data from iNaturalist was characterized by a number of data issues that result in inappropriate use of circular statistics (Forti, Hepp, et al., 2022). For analyses of diurnal activity, it is more likely that citizen scientists are making observations early in the night or just before sunrise, rather than during the middle of the night. Moreover, anurans are more likely to be observed (recorded or photographed) and uploaded to iNaturalist when engaging in reproductive behaviour (e.g. calling, coupling, laying eggs) than when silent or motionless. Regardless, it is essentially impossible to ensure equal effort in anuran observations per time interval throughout the night or year based on such data, compromising any ability to standardize effort or marginal totals for use in a bootstrap simulation. Within the context of phenological research, citizen science data are extremely valuable, but they are also systematically biased in ways that make statistical evaluations difficult. In general, citizen science data are collected when and where humans prefer to be, which can bias spatial and temporal estimates of activity or occurrence (Bird et al., 2014). This leads to unequal sampling effort and over-representation of observations in times and places that people choose to be in nature (e.g. Schubert et al., 2019; Forti, Pontes, et al., 2022). In addition, citizen scientists tend to over document and mis-identify rare species while simultaneously ignoring or under documenting common species (Bird et al., 2014). There is no apparent solution to account for such sampling biases within the context of circular statistics.

A study on dispersal of the Glossy Ibis (*Plegadis falcinellus*) in the Mediterranean Basin (Samraoui et al., 2023) represents an interesting variation on the problem of unequal effort among sampling intervals. Over 1000 fledglings from Numidian breeding colonies were banded. Thereafter, the locations of re-sightings were recorded over a 10-year period. The dispersal direction from breeding colonies was imputed from these locational data to evaluate if dispersal was uniform or directional. However, standardized sampling effort in all directions was not possible. Consequently, broad disparities in sampling effort characterized particular directions (and locations), including no sampling effort in some directions. Such haphazard and site-specific biases in detection have no obvious remedy when trying to evaluate directionality of fledgling dispersal.

### 4.5 | Recommendations for the future

Our simulation approach need not be limited to analyses based on Rayleigh or Hermans-Rasson tests. Indeed, analogous bootstrapping can be incorporated into analyses based on other statistical tests derived from transformations of $\alpha$ via sin and cosine functions (e.g. categorical analysis of variance, multivariate analysis of variance or periodic regression).

Importantly, the problem of marginal totals is not restricted to analyses of circular data via the Rayleigh test of Hermans-Rasson test. An innovative application of multivariate analysis of variance (MANOVA) to circular data (Landler et al., 2022) would also suffer from the same issues related to marginal totals, hidden marginal totals, and pseudoreplication as those found in analyses of circular data based on Rayleigh or Hermans-Rasson tests. However, application of our bootstrap in conjunction with the MANOVA would similarly rectify issues associated with marginal totals. Moreover, the use of multivariate analysis of covariance (MANCOVA) based on orthogonal transformations of $\alpha$ via sin and cosine functions as dependent variables, with marginal totals for each value of $\alpha$ as a covariate, may prove to be a flexible and powerful alternative approach worthy of future exploration. In short, a bootstrapping approach like the one we are presenting to standardize sampling effort or to account for data that are inherently binary (e.g. counting the number of

individuals that are reproductively active or reproductively inactive) is necessary prior to analysis using any type of circular method that does not inherently account for such data characteristics.

As with all analytical approaches, the sensitivity of our bootstrap is affected by multiple interacting factors, including sample size, distinctness of the pattern (e.g. activity peak height), and natural variation in responses to spatial, temporal, or environmental factors. For example, the sample size necessary to detect a distinctive pattern of reproductive activity is less than that required to detect a more subtle response. A more extensive set of sensitivity analyses may help to expose these complex interacting factors by explicitly considering variation in all combinations of them in a systematic fashion.

## 5 | CONCLUSIONS

Circular statistics is a critical complement to the arsenal of approaches used by ecologists to characterize temporal patterns in key biological activities. Nonetheless, some experimental designs or empirical data should not be analysed by such approaches. We identify substantive shortcomings in the use of circular statistics when the underlying empirical data are binomial (e.g. active versus non-active) and characterized by marginal totals (the sum of the number of active and non-active individuals per time interval). We then develop and apply a bootstrapping simulation approach to overcome these limitations, and illustrate its success with regard to exemplar data and empirical data on the reproductive phenology of tropical bats. In addition, two sets of sensitivity analyses demonstrate the ability of the bootstrap procedure to detect nonuniformity for different sample sizes and activity peak amplitudes. Finally, we further caution about the use of circular statistics in a number of different contexts characterized by pseudoreplication, marginal totals or biased sampling, which are common issues when trying to repurpose data for uses different than those associated with the original research. In some cases, alternative approaches can be used to test for uniformity in activity across intervals. In other cases, the design of the data collection may be too compromised to permit analysis by circular statistics.

### AUTHOR CONTRIBUTIONS

Michael R. Willig conceived and designed the overall study, gathered and provided the empirical data on bat phenologies, and wrote the first draft of the manuscript. Julissa Rojas-Sandoval designed, wrote the code, executed the bootstrapping simulation and crafted most of the figures. Steve J. Presley crafted the exemplar datasets and prepared the empirical data for analysis. All authors fully contributed to all aspects of the study, including the interpretation of results and writing text through multiple revisions.

### ACKNOWLEDGEMENTS

We thank D. Anglés-Alcázar for assistance and advice on the development of the bootstrapping approach as well as M. Prates for a review of and statistical comments on an earlier version of the manuscript. We are also grateful to J. Luna de Carvalho, who assisted with the collection of the empirical data on bat reproductive phenologies in Brazil. We thank the constructive reviews and suggestions from the reviewers and editor. M.R.W. and S.J.P. gratefully acknowledge support from the US National Science Foundation via an OPUS grant (DEB 1950643), and J.R.-S. is grateful for support from the College of Liberal Arts and Sciences and the Provost's Office at the University of Connecticut.

### CONFLICT OF INTEREST STATEMENT

The authors declare that there are no known competing financial interests and no personal relationships that have or will influence the research reported in this article. These codes can be used for assessing non-random phenologies when marginal totals characterize experimental designs.

### PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14316.

### DATA AVAILABILITY STATEMENT

Data used this study are available on Figshare: (https://doi.org/10.6084/m9.figshare.25146245.v3). Python codes used for the execution of the Rayleigh test and Hermans-Rasson test along with the code for bootstrapping simulations can be found on Zenodo: https://doi.org/10.5281/zenodo.10799004 (Willig et al., 2024).

### ORCID

*Michael R. Willig* https://orcid.org/0000-0001-6884-9957
*Julissa Rojas-Sandoval* https://orcid.org/0000-0001-6620-4741
*Steven J. Presley* https://orcid.org/0000-0002-5987-0735

### REFERENCES

Abe, T., Kubota, Y., Shimatani, K., Aakala, T., & Kuuluvainen, T. (2012). Circular distributions of fallen logs as an indicator of forest disturbance regimes. *Ecological Indicators*, *18*, 559–566.

Adrian, M. H., & Meeuwig, J. J. (2001). Detecting lunar cycles in marine ecology: periodic regression versus categorical ANOVA. *Marine Ecology Progress Series*, *214*, 307–310.

Alexiades, M. N. (1996). Standard techniques for collecting and preparing herbarium specimens. *Advances in Economic Botany*, *10*, 99–126.

Aradottir, A. L., Robertson, A., & Moore, E. (1997). Circular statistical analysis of birch colonization and the directional growth response of birch and black cottonwood in south Iceland. *Agricultural and Forest Meteorology*, *84*(1–2), 179–186.

Batschelet, E. (1981). *Circular statistics in biology*. Academic Press.

Beard, K. H., Kelsey, K. C., Leffler, A. J., & Welker, J. M. (2019). The missing angle: Ecosystem consequences of phenological mismatch. *Trends in Ecology & Evolution*, *34*(10), 885–888.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F., Pecl, G. T., Barrett, N., & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154.

Boullhesen, M., Vaira, M., Barquez, R. M., & Akmentins, M. S. (2023). Patterns of acoustic phenology in an anuran assemblage of the

Yungas Andean forests of Argentina. *Acta Herpetologica*, *18*, 23–36.

Calero, S., & Rodrigo, M. A. (2019). Reproductive phenology of submerged macrophytes: A tracker of year-to-year environmental variations. *Journal of Vegetation Science*, *30*(6), 1217–1227.

Castro-Arellano, I., Lacher, T. E., Jr., Willig, M. R., & Rangel, T. F. (2010). Assessment of assemblage-wide temporal niche segregation using null models. *Methods in Eecology and Evolution*, *1*(3), 311–318.

Castro-Arellano, I., Presley, S. J., Willig, M. R., Wunderle, J. M., & Saldanha, L. N. (2009). Reduced-impact logging and temporal activity of understorey bats in lowland Amazonia. *Biological Conservation*, *142*(10), 2131–2139.

Cooper, C. B., Shirk, J., & Zuckerberg, B. (2014). The invisible prevalence of citizen science in global research: Migratory birds and climate change. *PLoS One*, *9*, e106508.

Cortés-Flores, J., Andresen, E., Cornejo-Tenorio, G., & Ibarra-Manríquez, G. (2013). Fruiting phenology of seed dispersal syndromes in a Mexican neotropical temperate forest. *Forest Ecology and Management*, *289*, 445–454.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application (Cambridge series in statistical and probabilistic mathematics)*. Cambridge University Press.

De Valpine, P. (2023). The common sense of P values. *Ecology*, *95*, 617–621.

Durant, K. A., Hall, R. W., Cisneros, L. M., Hyland, R. M., & Willig, M. R. (2013). Reproductive phenologies of phyllostomid bats in Costa Rica. *Journal of Mammalogy*, *94*, 1438–1448.

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the bootstrap*. Chapman and Hall/CRC.

Eghbali, H., & Sharifi, M. (2023). Impacts of inter-annual climate variability on reproductive phenology and postnatal development of morphological features of three sympatric bat species. *Scientific Reports*, *13*(1), 8716.

Ellison, A. M., Gotelli, N. J., Inouye, B. D., & Strong, D. R. (2014). *P* values, hypothesis testing, and model selection: it's déjà vu all over again 1. *Ecology*, *95*(3), 609–610.

Estrada, A., & Coates-Estrada, R. (2001). Species composition and reproductive phenology of bats in a tropical landscape at Los Tuxtlas, Mexico. *Journal of Tropical Ecology*, *17*(5), 627–646.

Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge University Press.

Fitak, R. R., & Johnsen, S. (2017). Bringing the analysis of animal orientation data full circle: Model-based approaches with maximum likelihood. *Journal of Experimental Biology*, *220*(21), 3878–3882.

Forti, L. R., Hepp, F., de Souza, J. M., Protazio, A., & Szabo, J. K. (2022). Climate drives anuran breeding phenology in a continental perspective as revealed by citizen-collected data. *Diversity and Distributions*, *28*, 2094–2109.

Forti, L. R., Pontes, M. R., Augusto-Alves, G., Martins, A., Hepp, F., & Szabo, J. K. (2022). Data collected by citizen scientists reveal the role of climate and phylogeny on the frequency of shelter types used by frogs across the Americas. *Zoology*, *155*, 126052.

Garretson, A., & Forkner, R. E. (2021). Herbaria reveal herbivory and pathogen increases and shifts in senescence for northeastern United States maples over 150 years. *Frontiers in Forests and Global Change*, *4*, 664763.

Gomes, V. G. N., Valiente-Banuet, A., & Araujo, A. C. (2019). Reproductive phenology of cacti species in the Brazilian Chaco. *Journal of Arid Environments*, *161*, 85–93.

Gustafson, C. L., & Partch, C. L. (2015). Emerging models for the molecular basis of mammalian circadian timing. *Biochemistry*, *54*(2), 134–149.

Hazard, Q. C. K., Sabino-Pinto, J., Lopez-Naucells, A., Farnenda, F. Z., Meyer, C. F. J., & Rocha, R. (2022). Reproductive phenologies of phyllostomid bats in the Central Amazon. *Mammalian Biology*, *102*, 417–428.

Kharouba, H. M., Ehrlén, J., Gelman, A., Bolmgren, K., Allen, J. M., Travers, S. E., & Wolkovich, E. M. (2018). Global shifts in the phenological synchrony of species interactions over recent decades. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(20), 5211–5216.

Lakens, D. (2015). On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ*, *3*, e1142. https://doi.org/10.7717/peerj.1142

Landler, L., Ruxton, G. D., & Malkemper, E. P. (2018). Circular data in biology: Advice for effectively implementing statistical procedures. *Behavioral Ecology and Sociobiology*, *72*, 128.

Landler, L., Ruxton, G. D., & Malkemper, E. P. (2019). The Hermans-Rasson test as a powerful alternative to the Rayleigh test for circular statistics in biology. *BMC Ecology*, *19*, 30.

Landler, L., Ruxton, G. D., & Malkemper, E. P. (2022). The multivariate analysis of variance as a powerful approach for circular data. *Movement Ecology*, *10*, 21.

Lieth, H. (1974). *Purposes of a phenology book* (pp. 3–19). Springer.

Lima, D. F., Mello, J. H. F., Lopes, I. T., Forzza, R. C., Goldenberg, R., & Freitas, L. (2021). Phenological responses to climate change based on a hundred years of herbarium collections of tropical Melastomataceae. *PLoS One*, *16*, e0251360.

Lima, T. N., Morellato, L. P. C., & Machado, I. C. (2012). Reproductive phenology of a northeast Brazilian mangrove community: Environmental and biotic constraints. *Flora-Morphology, Distribution, Functional Ecology of Plants*, *207*(9), 682–692.

Lorch, P. D., Sword, G. A., Gwynne, D. T., & Anderson, G. L. (2005). Radiotelemetry reveals differences in individual movement patterns between outbreak and non-outbreak Mormon cricket populations. *Ecological Entomology*, *30*, 548–555.

Lovari, S., Corsini, M. T., Guazzini, B., Romeo, G., & Mori, E. (2017). Suburban ecology of the crested porcupine in a heavily poached area: A global approach. *European Journal of Wildlife Research*, *63*, 10.

Morellato, L. P. C., Alberti, L. F., & Hudson, I. L. (2010). *Applications of crcular statistics in plant phenology: A case studies approach* (pp. 339–359). Springer.

Muff, S., Nilsen, E. B., O'Hara, R. B., & Nater, C. R. (2022). Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution*, *37*, 203–210.

Nurul-Ain, E., Rosli, H., & Kingston, T. (2017). Resource availability and roosting ecology shape reproductive phenology of rain forest insectivorous bats. *Biotropica*, *49*(3), 382–394.

Ossi, F., Ranc, N., Moorcroft, P., Bonanni, P., & Cagnacci, F. (2020). Ecological and behavioral drivers of supplemental feeding use by roe deer *Capreolus capreolus* in a peri-urban context. *Animals*, *10*, 2088.

Ożarowska, A., Ilieva, M., Zehtindjiev, P., Åkesson, S., & Muś, K. (2013). A new approach to evaluate multimodal orientation behaviour of migratory passerine birds recorded in circular orientation cages. *Journal of Experimental Biology*, *216*(21), 4038–4046.

Pabon-Moreno, D. E., Musavi, T., Migliavacca, M., Reichstein, M., Römermann, C., & Mahecha, M. D. (2020). Ecosystem physiophenology revealed using circular statistics. *Biogeosciences*, *17*(15), 3991–4006.

Rother, D. C., de Sousa, I. L. F., Gressler, E., Liboni, A. P., Souza, V. C., Rodrigues, R. R., & Morellato, L. P. C. (2022). Comparing the potential reproductive phenology between restored areas and native tropical forest fragments in Southeastern Brazil. *Restoration Ecology*, *30*, e13529.

Samraoui, B., Nedjah, R., Boucheker, A., Bouzid, A., El-Serehy, H. A., & Samraoui, F. (2023). Blowin' in the wind: Dispersal of glossy ibis *Plegadis falcinellus* in the West Mediterranean basin. *Ecology and Evolution*, *13*, e9756.

Scheiner, S. M., & Willig, M. R. (Eds.). (2011). *Theory of ecology*. University of Chicago Press.

Schubert, S. C., Manica, L. T., & De Camargo Guaraldo, A. (2019). Revealing the potential of a huge citizen-science platform to study bird migration. *Emu - Austral Ornithology*, *119*, 364–373.

Sheldon, K. S., & Nadkarni, N. M. (2015). Reproductive phenology of epiphytes in Monteverde, Costa Rica. *Revista de Biología Tropical*, *63*(4), 1119–1126.

Shimatani, I. K., Yoda, K., Katsumata, N., & Sato, K. (2012). Toward the quantification of a conceptual framework for movement ecology using circular statistical modeling. *PLoS One*, *7*(11), e50309.

Staggemeier, V. G., Camargo, M. G. G., Diniz-Filho, J. A. F., Freckleton, R., Jardim, L., & Morellato, L. P. C. (2020). The circular nature of recurrent life cycle events: A test comparing tropical and temperate phenology. *Journal of Ecology*, *108*(2), 393–404.

Taube, J. S. (2007). The head direction signal: Origins and sensory-motor integration. *Annual Review of Neuroscience*, *30*, 181–207.

Upton, G. J. G., & Fingleton, B. (1989). *Spatial data analysis by example*. Volume 2. Categorical and Directional Data. John Wiley & Sons.

Valentin-Silva, A., Staggemeier, V. G., Batalha, M. A., & Guimarães, E. (2018). What factors can influence the reproductive phenology of Neotropical Piper species (Piperaceae) in a semi-deciduous seasonal forest? *Botany*, *96*(10), 675–684.

Vallat, R. (2018). Pingouin: Statistics in python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

Wicker, E. (2021). Bootstrap resampling. An overview and example with scikit-learn's resample and BaggingRegressor. https://ethanwicker.com/2021-02-23-bootstrap-resampling-001/

Willig, M., Rojas-Sandoval, J., & Presley, S. (2024). Phenological patterns in ecology: Problems using circular statistics and solutions based on simulations. *Zenodo*. https://doi.org/10.5281/zenodo.10799004

Willig, M. R. (1985a). Ecology, reproductive biology, and systematics of *Neoplatymops mattogrossensis* (Chiroptera: Molossidae). *Journal of Mammalogy*, *66*, 618–628.

Willig, M. R. (1985b). Reproductive patterns of bats from Caatingas and Cerrado biomes in northeast Brazil. *Journal of Mammalogy*, *66*, 668–681.

Willig, M. R., & Presley, S. J. (2023). Reproductive phenologies of bat populations and ensembles from lowland Amazonia. *Journal of Mammalogy*, *104*, 752–769.

Wilson, D. E. (1973). Reproductive patterns. In R. J. Baker, J. K. Jones, Jr., & D. C. Carter (Eds.), *Biology of bats of the new world family Phyllostomatidae* (pp. 317–378). Special Publications, Texas Tech University.

Wiltschko, W., & Wiltschko, R. (1972). Magnetic compass of European robins. *Science*, *176*(4030), 62–64.

Wolkovich, E. M., Cook, B. I., McLauchlan, K. K., & Davies, T. J. (2014). Temporal ecology in the Anthropocene. *Ecology Letters*, *17*(11), 1365–1379.

Zar, J. H. (2009). *Biostatistical analysis* (5th ed.). Prentice-Hall.

**How to cite this article:** Willig, M. R., Rojas-Sandoval, J., & Presley, S. J. (2024). Phenological patterns in ecology: Problems using circular statistics and solutions based on simulations. *Methods in Ecology and Evolution*, *00*, 1–18. https://doi.org/10.1111/2041-210X.14316