Chapter 4
Print ISBN: 978-93-47485-65-7, eBook ISBN: 978-93-47485-07-7

Statistical Techniques in Ecology: Descriptive Statistics and Normal Distribution

B. K. Singh [a], Rajan Singh [a*], Anshul Dubey [a], Nidhi Tiwari [a] and Nidhi Prabhakar [a]

DOI: https://doi.org/10.9734/bpi/mcsru/v9/7057

Peer-Review History:
This chapter was reviewed by following the Advanced Open Peer Review policy. This chapter was thoroughly checked to prevent plagiarism. As per editorial policy, a minimum of two peer-reviewers reviewed the manuscript. After review and revision of the manuscript, the Book Editor approved the manuscript for final publication. Peer review comments, comments of the editor(s), etc. are available here: https://peerreviewarchive.com/review-history/7057

Abstract

Ecological science relies on robust estimates of the abundance, diversity, and spatial distribution of individuals and species, but these quantities are notoriously difficult to observe directly. Statistics may be considered as the science and technique of collecting, analysing, and making inferences from data, and these references are stated as probabilities. The study aims to explore and apply quantitative and statistical methods in ecology to understand the relationships between populations and their environment, assess the effects of environmental hazards on animal and plant populations, and evaluate overall ecological balance. Fundamental statistical concepts, including descriptive statistics, probability distribution, regression and correlations, and the chi-square distribution, are demonstrated to show their function in analysing ecological data. On the other hand, specialised methods, such as species-abundance relations and species-diversity measures, provide insights into community structure and ecosystem stability. The study recommends the use of logarithmic distributions to accurately fit species-abundance data and enhance the reliability of ecological analyses.

Keywords: Environment; biomass of a population; species-abundance relations; mean and variance; standard deviation; histogram.

1 Introduction

The study of ecology has become quantitative and mathematical during the last few decades because of the general awareness of the importance of the relationships between the environment and biomass of a population, interactions between populations and within populations and the effect of population on the environment (Amrhein and Greenland, 2022). Different kinds of pollution (e.g. water, air, noise, etc.) have appeared as a real threat to the existence of the human population, the most intelligent species on

[a] *Department of Mathematics, School of Sciences, IFTM University, Moradabad-244102, Uttar Pradesh, India.*
**Corresponding author: E-mail: rajansingh@iftmuniversity.ac.in;*

planet earth (Altwegg et al., 2025). To measure the effect of these hazards on the animal and plant populations and on the overall ecological balance, the knowledge of different statistical methods and techniques has become indispensable. In the first part of this chapter, elementary concepts of descriptive statistics, probability distribution, regression and correlations, the chi-square distribution, etc., will be reviewed briefly, while in the second part we shall dwell on the specialised topic, namely species-abundance relations, and the measurement of species-diversity.

2 Concepts of Statistical Techniques

Statistics may be considered as the science and technique of collecting, analysing, and making inferences from data, and these references are stated as probabilities (Crowe and Cash, 2023). Three flawed practices associated with model averaging coefficients for predictor variables in regression models commonly occur when making multimodal inferences in analyses of ecological data (Cade, 2015). The term probability is used in the sense of the relative frequency of multiple or repeated events in the long run (Janas et al., 2020). The data under consideration for statistical analysis consists of two kinds of variables: (a) continuous and (b) discrete (Berry et al., 2021). The former consists of measurements against a suitable scale, while the latter consists of the data obtained by counting or enumerating discrete, indivisible units (Dubey and Singh, 2020). Mean and Variance are the best-known statistical measures of a series of observations. Suppose the following numbers of individuals have been observed in a series of 10 quadrats:

16, 11, 19, 28, 34, 62, 18, 20, 10, 12

The average number of individuals in a quadrat, the mean, is the sum of all observations divided by the number of observations

$$\bar{x}\,(\text{mean}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

So, $\bar{x} = (16 + 11 + 19 + 28 + 34 + 62 + 18 + 20 + 10 + 12)/10 = 23$ individuals. To have an insight, it is desirable to estimate the average deviation of each observed value from the mean$\bar{x}$. The average deviation measured from the mean is always zero because the sum of the deviations is zero. To eliminate this problem, the squared deviation is usually calculated, i.e. $\sum_i (x_i - \bar{x})^2$. The average of the squared deviations is called variance σ^2 where σ is called the standard deviation. Therefore

$$\sigma^2 = \frac{1}{n}\sum_{i} (x_i - \bar{x})^2 = \left[\frac{\sum {x_i}^2}{n} - \left(\frac{\sum x_i}{n}\right)^2\right]$$

Table 1. Sum of 10 quadrats and their square values

x_i	x_i^2
16	256
11	121

x_i	x_i^2
19	361
28	784
34	1156
62	3844
18	324
20	400
10	100
12	144
$\sum x_i = 230$	$\sum x_i^2 = 7490$

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$
$$= [\frac{\sum x_i^2}{n} - (\frac{\sum x_i}{n})^2]$$
$$= [\frac{7490}{10} - (\frac{230}{10})^2]$$
$$= 749 - 529$$
$$= 220\ \sigma^2 = 220$$
$$\sigma = \pm 14.83$$

In the present example $\sigma^2 = 220$, and$\sigma = \pm 14.83$.

Let us now consider the weights of 2,089 individuals in a population shown in Table 2. The observations are divided into 10-kilogram groupings. A total of 216 individuals were found to weigh between 70 and 80 kilograms.

Table 2. The Frequency Distribution of the Individuals

Weight Group	No. of Observation	Fraction of Total
30-40	13	0.03714
40-50	30	0.06666
50-60	68	0.12364
60-70	135	0.20769
70-80	216	0.288
80-90	300	0.35294
90-100	345	0.36316
100-110	338	0.32190
110-120	275	0.23913
120-130	190	0.152
130-140	101	0.07481
140-150	48	0.03310
150-160	20	0.01290
160-170	7	0.00424
170-180	3	0.00171
	N=2089	

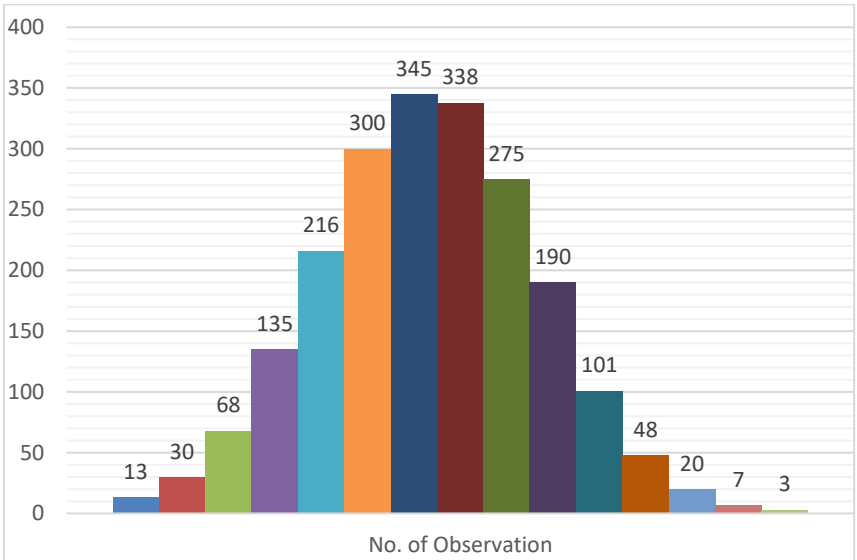

Fig. 1a. Histogram of the frequency distribution of the individuals

The number of observations occurring in each group is plotted against the groups. This is known as frequency distribution. From the Histogram, it is observed that it is approximately bell-shaped, most of the observations are centred towards the middle groups, and a few are at the two tails of the distribution (Dixon et al., 2005).

If the length of each class interval becomes smaller and smaller, the discontinuities of the distribution become smaller and smaller. Considering the weight groups to be infinitesimally small, the distribution would resemble the line curve as shown in Fig. 1.b. In Fig. 2, the dotted curve is shown clearly. It is a bell-shaped curve known as the normal curve.

This curve represents a function known as the normal probability density function. The density function is continuous, and the distribution in Fig. 1 approaches the normal curve in shape in the limit as the class interval becomes smaller and smaller.

In the previous discussion, the idea of the normal or Gaussian curve was introduced, which is taken as the basis of most of the applications of statistics in biology, physiology, ecology, etc. The curve is also known as the error law, and historically, the equation for the normal curve was based on the analysis of the distribution of errors or deviations around the mean of the physical measurements by celebrated mathematician Carl F (Ellison and Dennis, 2010). Gauss. In analysing a set of sampled data, it is assumed that they could fit the theoretical distribution of the normal curve closely. The reason for this closeness of the approximation of the theoretical normal curve to real data lies in the random interaction of many small and variable factors (DiRenzo et al., 2023). This does not necessarily imply that one can take the normal distribution in the analysis of data

merely because the factors under consideration are many and small. Often, it is observed that a particular observation is a nonlinear function of some variable which is normally distributed.

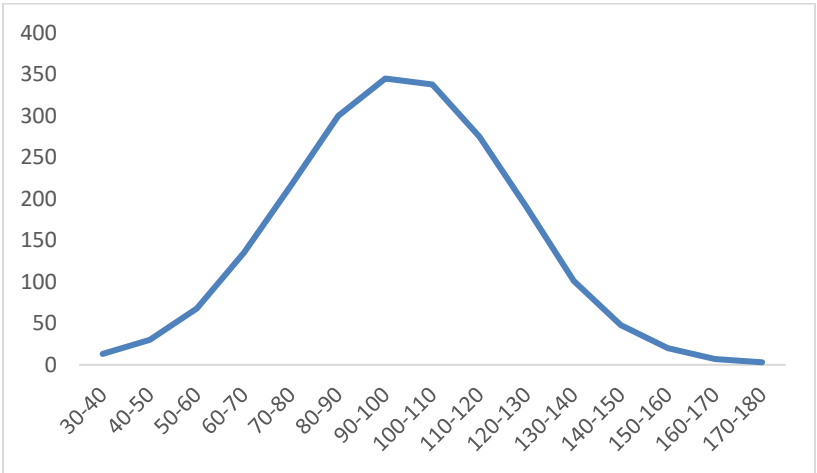

Fig. 1b. Curve showing the Frequency Distribution of the Individuals

The Gaussian or normal curve is described by the equation

$$y = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{(x-\mu)^2}{2\sigma^2}]$$

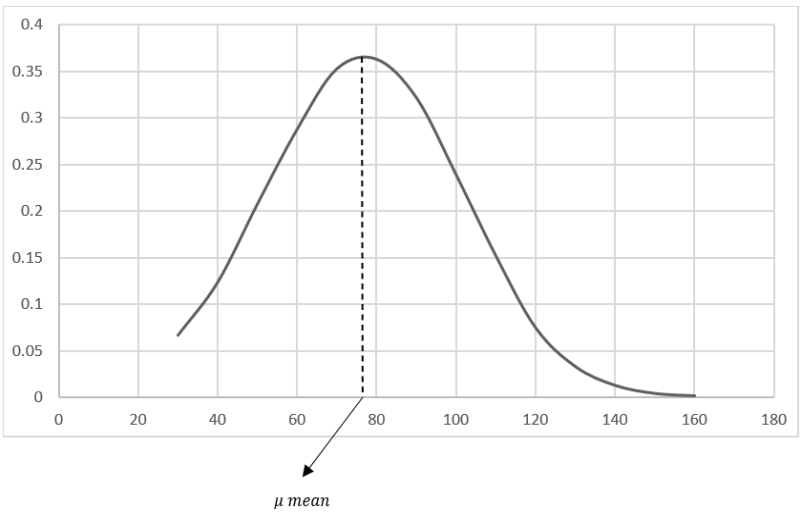

Fig. 2. Normal probability density function

In this equation, y represents the relative frequency of some variable quantity x. The values for transcendental numbers π and e are constant (Gardner et al., 2022). This equation has two important parameters μ, the arithmetic means and σ, the standard deviation, which may be taken as the measure of the spread of the data about the mean (Ellison and Dennis, 2010). The curve is completely determined if the values of the parameters μ and σ are known, since π and e are constants. Knowledge of the following important properties of the normal curve is important:

(a) It is a bell-shaped, symmetrical curve. The median and mode coincide with mean μ.
(b) It is initially convex upward but soon becomes concave. There is a point of transition from convex to concave called a point of inflexion. The distance of this point horizontally from the mean is equal to the standard deviation σ.
(c) The curve tapers out to infinity in either direction, from the mean before approaching the horizontal axis (x-axis).
(d) The area under the curve is unity, i.e. the sum of all the probabilities (possible relative frequencies) represents certainty.

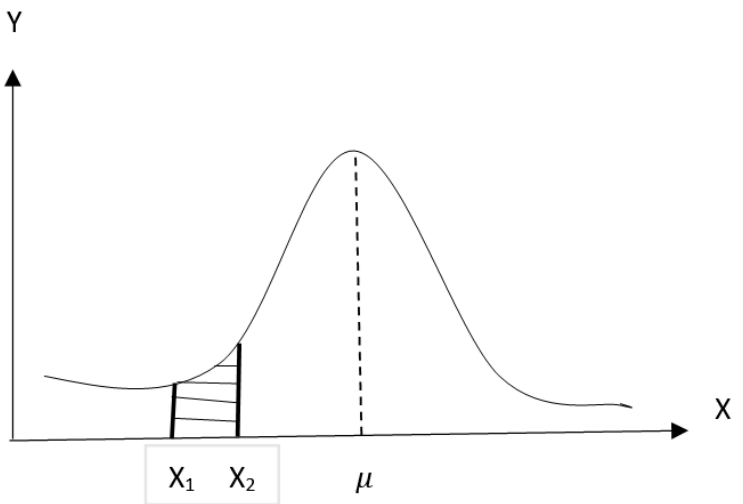

Fig. 3. Normal curve showing area proportional to probability

Thus, the area lying under the curve between the values x_1 and $x_2$shown as the shaded area in Fig. 3, represents a fraction of the total are proportion to its probability. In other words, this fraction represents the probability of obtaining a sample value lying between x_1 and x_2. As the curve is symmetrical, half the area under the curve lies above the mean, while half lies below it.

The shaded area shown in Fig. 4 is the area lying under the curve between the inflexion points. This area constitutes about 2/3 of the total area, or more-approximately 67%.

(e) The equation of the normal curve can be expressed in the form

$$y = \frac{1}{\sqrt{2\pi}} e^{-z}$$

Where

$$Z = \frac{x-\mu}{\sigma}$$

This is known as the normal curve in the standard form. Here mean is zero and the standard deviation is unity. The spreading of a normal curve about the mean is ± 3 and in the standard form, it is ± 3 about the mean zero.

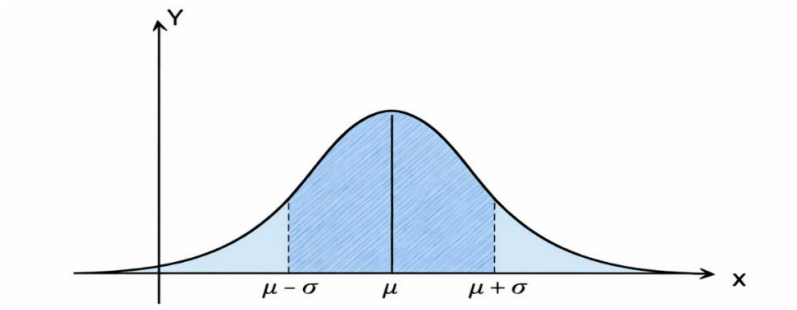

Fig. 4. Area between the inflexion points

The discussion so far is based on the assumption that the estimates of the mean and standard deviations are accurately known because of the large sample sizes. But soon we recognise that the parameters of any universe of discourse are never known exactly (Ellison and Dennis, 2010). In general, the mean and standard deviation of any universe of discourse are unknowable values, but one can arrive at a satisfactory evaluation of the magnitude of error in an estimate by considering the effect of sample size on the mean (Dubey and Singh, 2022).

In Fig. 4, the curve for $n = 1$ represents the normal distribution curve of a population whose mean is μ and standard deviation (S.D.) is σ. Suppose that in estimating the characteristics of a population, instead of making individual measurements, we make the measurements in pairs and treat the means of the two measurements as a variate. This derived population has a narrower scatter, but the mean will be identical to the universe mean of the parent population. These distributions are shown in Fig. 5. If we increase the size of the samples by making the subgroups, the means remain unchanged, but S.d. of the distribution of sample means diminishes. If σ is the standard deviation of a universe, it can be shown that the standard deviation of the means for samples of n variates is $\sigma/\sqrt{n}$ and the value is called the standard error (S.E.) of the mean; 1.4 is still a normal distribution. Expressing the proportionate area as a function of standardized deviate *Z*, we get

$$Z = \pm \frac{(\mu - \bar{x})}{s/\sqrt{n}}$$

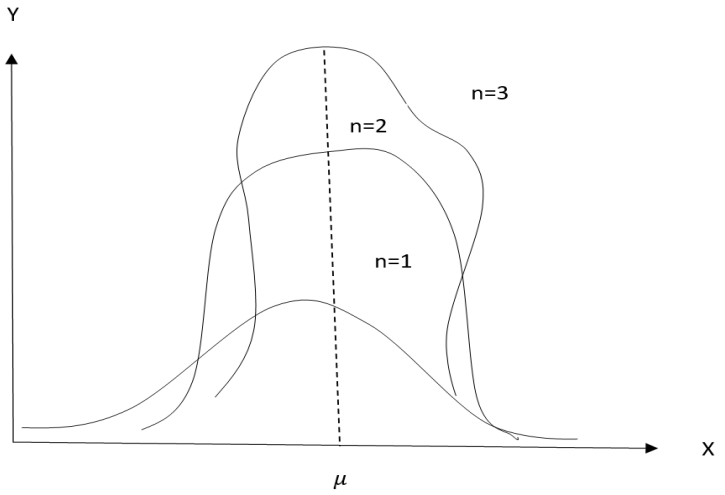

Fig. 5. Effect of sample size on the distribution of sample mean

Where, $\bar{x}$ is the measured estimate for the mean, and *s* is the standard deviation.

By transformation, we can write

$$\bar{x} = \mu \pm \frac{Zs}{\sqrt{n}}, \text{ or } \mu = \bar{x} \mp \frac{Zs}{\sqrt{n}}$$

The above equation states that the estimate of the mean using a sample size n lies within a certain range of the universe mean μ, if we can select an appropriate value for *Z (Dubey et al., 2024)*.

Suppose we select a value of *Z* for which the excluded or shaded area constitutes 5% of the total area (Singh, 2008). This means that 95% of the estimates, $\bar{x}$ will lie no further from μ than $Zs/\sqrt{n}$. This value will be found to be 1.96. In other words

$$\mu - 1.96\frac{s}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{s}{\sqrt{n}}$$

As an example, let us consider the mean weight of all men in the world to be 130lb and the standard deviation to be 8 lb. Then, 95% of the time, the mean weights of samples of 100 women selected at random would fall within the limits described by the equation

$$\bar{x} = 130 \pm 1.96(\frac{8}{\sqrt{100}})$$

In biological and ecological studies, 95% probability is conventionally accepted as adequate assurance. If the limits are extended from 1.96s to 2*s* on either side of the mean to 3*s*, the probability goes up from 95% upto 99%. Thus, for a 50% increase in range

around the mean, we gain a mere 4% increased assurance. Conventionally, we speak of the limits around the mean $\pm 1.96s/\sqrt{n}$, as the 95% confidence limits or fiducial limits. Sometimes, in real situations, it is not always feasible to have a large sample (Lájer, 2007). With small samples, we cannot do this safely without making some allowance for the unreliability of s as a measure of σ. We do this in the following way by multiplying Z by a factor g while setting limits of the mean fiducially, namely

$$\mu = \bar{x} \pm (gZ)s$$

Where g depends on the sample size n as $g(n) > 1$ when n is small, but approaches the value 1 as n increases. The correction factor gZ was solved first by W.S. Gesset (1908) under the pseudonym 'Student'. The various values of $g(n)$ for different sample size n were tabulated under the heading 't' and the use of this corrected value is known as Student`s t-test. For example, n=30 or more, t and Z are practically equal. Therefore

$$\mu = \bar{x} \pm ts, \text{ where } n < 30$$

3 Abundance and Diversity of Species

Many species of organisms are observed in most ecological communities and these species vary greatly in their abundance from very common to very rare (Bonet and Pausas, 2004). One of the most important areas of investigation now–a-days to the ecologist is to ascertain the distribution of abundances of the different species and from that to have the knowledge of rare species (Lüdecke, 2021). Due to different types of environmental pollution, different species are adversely affected, and the existence of certain species is threatened (Ioannidis, 2018). To protect the rare species from extinction, the distribution pattern of abundance is essential. Different mathematical distributions for the species-abundance relationship will be considered for different measures of diversity of species, particularly from the viewpoint of information theory (Kardish et al., 2015). Insect counts in the field (and other population counts) are often fitted fairly well by a negative binomial distribution (Bliss, 1958). Over the last decade, spatial capture–recapture (SCR) models have become widespread for estimating demographic parameters in ecological studies (Gardner et al., 2022).

The simplest way to assess species diversity is by counting how many species from a given taxonomic group occur in an area; however, the relative abundance of each species is also an important factor (Dubey and Singh, 2019). Numerous approaches have been developed to apply various mathematical models to describe species–abundance relationships. The usual practice of calculating the arithmetic mean in a set of measurements of a biological nature (Fisher et al., 1943). The goal is to identify a model that accurately represents data from a wide range of communities and enables meaningful comparisons among them based on the parameters of the distribution (Fleming et al., 2015). One approach is the listing of species-abundance data. Instead of listing the number of individuals in species 1, species 2, etc., the number of species n_1, represented by one member, the number of species n_r, represented by r members, and so on, are listed. The symbol n_r represented frequency of species with r individuals. The same numerical predictions of the average relative abundance of species that follow from MacArthur's (1957) "broken stick" model also follow from a "balls and boxes" model

(Cohen, 1966) with a different set of assumptions (Cohen, 1968). In many scientific disciplines, common research practices have led to unreliable and exaggerated evidence about scientific phenomena (Kimmel et al., 2023).

4 Conclusion

Ecology has evolved into a highly quantitative field of study due to the growing need to understand complex interactions between organisms, their populations, and the environment under increasing human-induced stress. A study of basic statistical concepts, including probability distributions, Mean and Variance, Standard Deviation, regression, and descriptive statistics, is necessary for the analysis of ecological data. These techniques help identify patterns and quantify relationships within and between populations. Building on this foundation, studies of species-abundance correlations and species diversity measures offer a deeper comprehension of community structure, ecosystem stability, and the consequences of environmental change. In ecology, the true causal structure for a given problem is often not known, and several plausible models and thus model predictions exist (Dormann et al., 2018).

5 Recommendation

We shall now consider some of the distributions which have been proposed to fit the observed species-abundance frequency distribution. The logarithmic distribution will be discussed first, because it is useful in providing an empirical fit to the observed species-abundance relationship.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during the writing or editing of this manuscript.

Competing Interests

Authors have declared that no competing interests exist.

References

Altwegg, R., Salau, S., Abadi, F., Cervantes, F., Clark, A. E., Distiller, G., ... & Visser, V. (2025). Emerging topics and new directions in statistical ecology. *Journal of Statistical Theory and Practice*, *19*(3), 44. https://doi.org/10.1007/s42519-025-00460-4

Amrhein, V., & Greenland, S. (2022). Rewriting results in the language of compatibility. *Trends in Ecology & Evolution, 37*(7), 567–568. https://doi.org/10.1016/j.tree.2022.02.001

Berry, O., Jarman, S., Bissett, A., Hope, M., Paeper, C., Bessey, C., Schwartz, M. K., Hale, J., & Bunce, M. (2021). Making environmental DNA (eDNA) biodiversity records globally accessible. *Environmental DNA, 3*, 699–705. https://doi.org/10.1002/edn3.173

Bliss, C. I. (1958). The analysis of insect counts as negative binomial distributions. *Proceedings of the X International Congress of Entomology, 2*, 1015–1032.

Bonet, A., & Pausas, J. G. (2004). Species richness and cover along a 60-year chronosequence in old-fields of southeastern Spain. *Plant Ecology, 174*, 257–270. https://doi.org/10.1023/B:VEGE.0000049106.96330.9c

Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology, 96*, 2370–2382. https://doi.org/10.1890/14-1639.1

Cohen, J. E. (1968). Alternate derivations of species abundance relation. *The American Naturalist, 102*, 165–172. https://doi.org/10.1086/282533

Crowe, R. P., & Cash, R. E. (2023). A letter from the editors: Pearls and pitfalls for writing a methods section. *Prehospital Emergency Care*, 27(2), 1–6. https://doi.org/10.1080/10903127.2023.2166177

DiRenzo, G. V., Hanks, E., & Miller, D. A. W. (2023). A practical guide to understanding and validating complex models using data simulations. *Methods in Ecology and Evolution, 14*, 203–217. https://doi.org/10.1111/2041-210X.14030

Dixon, P. M., Ellison, A. M., & Gotelli, N. J. (2005). Improving the precision of estimates of the frequency of rare events. *Ecology, 86*, 1114–1123.

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Barton, K., Beale, C. M., et al. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs, 88*, 485–504. https://doi.org/10.1002/ecm.1309

Dubey, A., & Singh, R. (2019). The effects of insulin in diabetes and statistical analysis through the correlation theory. *Journal of Emerging Technologies and Innovative Research (JETIR), 6*(6), 393–396. ISSN 2349-5162. https://www.jetir.org/view?paper=JETIR1907P57

Dubey, A., & Singh, R. (2020). The statistical analysis through the Newton's divided difference interpolation in diabetic patients. *International Journal of Research and Analytical Reviews (IJRAR), 7*(1), 833–837. E-ISSN 2348-1269, P-ISSN 2349-5138. https://www.ijrar.org/IJRAR2002114.pdf

Dongre, P., & Verma, P. (2022). Securing IoT with Blockchain-Based System for Attendance Management. Stochastic Modeling & Applications, 26(3), 1073–1080. https://mukpublications.com/journals/stochastic-modeling-applications/vol-26-no-3-2022/

Dubey, A., Singh, R., Singh, B. K., & Tiwari, N. (2024). Statistical analysis through the chi-square test in the reported cases of malaria at Moradabad in Uttar Pradesh. *Global Journal of Engineering and Technology (GJET), 3*(2), 34–36. ISSN 2583-3359 (Online). https://gsarpublishers.com/gjet-vol-3-issue-2-february-2024/

Ellison, A. M., & Dennis, B. (2010). Paths to statistical fluency for ecologists. *Frontiers in Ecology and the Environment, 8*, 362–370. https://doi.org/10.1890/080209

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology, 12, 42–58. https://doi.org/10.2307/1411

Fleming, C. H., Fagan, W. F., Mueller, T., Olson, K. A., Leimgruber, P., & Calabrese, J. M. (2015). Rigorous home range estimation with movement data: A new autocorrelated kernel density estimator. *Ecology, 96*, 1182–1188. https://doi.org/10.1890/14-2010.1

Gardner, B., McClintock, B. T., Converse, S. J., & Hostetter, N. J. (2022). Integrated animal movement and spatial capture–recapture models: Simulation, implementation, and inference. *Ecology, 103*, e3771. https://doi.org/10.1890/14-2010.1

Ioannidis, J. P. A. (2018). The proposal to lower P value thresholds to .005. *JAMA, 319*(14), 1429–1430. https://doi.org/10.1001/jama.2018.1536

Janas, K., Lutyk, D., Sudyka, J., Dubiec, A., Gustafsson, L., Cichoń, M., & Drobniak, S. (2020). Carotenoid-based coloration correlates with the hatching date of blue tit *Cyanistes caeruleus* nestlings. *Ibis, 162*(3), 645–654. https://doi.org/10.1111/ibi.12751

Kardish, M. R., Mueller, U. G., Amador-Vargas, S., Dietrich, E. I., Ma, R., Barrett, B., & Fang, C.-C. (2015). Blind trust in unblinded observation in ecology, evolution, and behavior. *Frontiers in Ecology and Evolution*. https://doi.org/10.3389/fevo.2015.00051

Kimmel, K., Avolio, M. L., & Ferraro, P. J. (2023). Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nature Ecology & Evolution, 7*(9), 1525–1536. https://doi.org/10.1038/s41559-023-02144-3

Lájer, K. (2007). Statistical tests as inappropriate tools for data analysis performed on non-random samples of plant communities. *Folia Geobotanica*, 42, 115–122. https://www.ibot.cas.cz/folia/42/2/lajer.pdf

Lüdecke, D. (2021). *sjPlot: Data visualization for statistics in social science* (R package version 2.8.10). https://doi.org/10.32614/CRAN.package.sjPlot

Singh, P. (2008). *Modeling crop production systems: Principles and application*. CRC Press. https://doi.org/10.1201/9781482280449

Peer-Review History:
This chapter was reviewed by following the Advanced Open Peer Review policy. This chapter was thoroughly checked to prevent plagiarism. As per editorial policy, a minimum of two peer-reviewers reviewed the manuscript. After review and revision of the manuscript, the Book Editor approved the manuscript for final publication. Peer review comments, comments of the editor(s), etc. are available here: https://peerreviewarchive.com/review-history/7057