Marques, T.A., S.T. Buckland, D.L. Borchers, E. Rexstad and L. Thomas. 2011. Distance sampling. Pp 398-400 in Lovric, M. (Ed.), International Encyclopedia of Statistical Science, Springer-Verlag, Berlin.
DOI of article: 10.1007/978-3-642-04898-2_214

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are two multivariate observations, Σ^{-1} is the inverse of the variance-covariance matrix and $(\boldsymbol{a}-\boldsymbol{b})'$ is the transpose of vector $(\boldsymbol{a}-\boldsymbol{b})$.

The Mahalanobis distance is designed to take into account the correlation between all variables (attributes) of the observations under consideration. For uncorrelated variables, the Mahalanobis distance reduces to the Euclidean distance for standardized data.

As an example, consider a set of points $\boldsymbol{x}$ in R^2 that have the constant distance r from the origin, that is, $(0,0)$. Then, the set of points having the property $d^2_{Mahalanobis}(0,\boldsymbol{x}) = r$ is an ellipse. The Mahalanobis distance is a positive definite quadratic form $\boldsymbol{x}'\mathbf{A}\boldsymbol{x}$, where the matrix $\mathbf{A} = \Sigma^{-1}$.

Distance measures or metrics are members of a broader concept called *similarity measures* (or *dissimilarity measures*) (Theodoridis and Koutroumbas 2009) that measure likeness (or affinity) in the case of the similarity measure, or difference (or lack of affinity) in the case of dissimilarity between objects. Similarity measures can be converted to dissimilarity measures using a monotone decreasing transformation and vice versa.

The main difference between metrics and broader concepts of similarity/dissimilarity measures is that some of the properties (1)–(4) do not hold for similarity/dissimilarity measures. For example, definiteness, or the triangle inequality, usually do not hold for similarity/dissimilarity measures.

The cosine similarity and the Pearson's product moment coefficient are two similarity measures that are not metric. The cosine similarity is the cosine of an angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ from R^n and is given by:

$$s(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|},$$

where $\|\boldsymbol{x}\|$ and $\|\boldsymbol{y}\|$ are norms of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. This measure is very popular in information retrieval and text-mining applications.

In statistical analysis (especially when applied to ecology, natural language processing, social sciences, etc.) there are often cases in which similarity or the distance between two items (e.g., sets, binary vectors) is based on two-way contingency tables with elements a, b, c, and d, where a represents the number of elements (attribute values, variables values) present in both items, b is the number of elements present in the first but absent in the second item, c is the number of elements present in the second but absent in the first item, and d is number of elements absent simultaneously in both items. The numbers a, b, c, and d can be defined as properties of two sets or two binary vectors.

Similarity coefficients (Theodoridis and Koutroumbas 2009) or *associations measures* can be defined as a combination of numbers a, b, c, and d. Examples of associations measures are:

Simple matching coefficient	$(a+d)/n$,
Dice coefficient	$2a/(2a+b+c)$,
Jaccard (or *Tanimoto*) *coefficient*	$a/(a+b+c)$.

Although association measures, similarity measures, and correlation coefficients are not metric, they are applicable in the analysis where they are consistent with the objective of the study and where they have meaningful interpretation (Sharma 1996).

Cross References

►Cook's Distance
►Data Mining Time Series Data
►Entropy and Cross Entropy as Diversity and Distance Measures
►Multidimensional Scaling: An Introduction
►Multivariate Outliers
►Statistical Natural Language Processing
►Statistics on Ranked Lists

References and Further Reading

Mardia KV, Kent JT, Bibby JM (1982) Multivariate analysis. Academic, New York
Rudin W (1976) Principles of mathematical analysis, 3rd edn. McGraw Hill, New York
Sharma SC (1996) Applied multivariate techniques. Wiley, New York
Theodoridis S, Koutroumbas K (2009) Pattern recognition, 4th edn. Elsevier

Distance Sampling

Tiago A. Marques[1,2], Stephen T. Buckland[1], David L. Borchers[1], Eric A. Rexstad[1], Len Thomas[1]
[1]University of St Andrews, St Andrews, UK
[2]Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal

Distance sampling is a widely used methodology for estimating animal density or abundance. Its name derives from the fact that the information used for inference are the recorded distances to objects of interest, usually animals, obtained by surveying lines or points. The methods

are also particularly suited to plants or immotile objects, as the assumptions involved (see below for details) are more easily met. In the case of lines the perpendicular distances to detected animals are recorded, while in the case of points the radial distances from the point to detected animals are recorded. A key underlying concept is the detection function, usually denoted $g(y)$ (here y represents either a perpendicular distance from the line or a radial distance from the point). This represents the probability of detecting an animal of interest, given that it is at a distance y from the transect. This function is closely related to the probability density function (pdf) of the detected distances, $f(y)$, as

$$f(y) = \frac{g(y)\pi(y)}{\int_0^w g(y)\pi(y)dy}, \quad (1)$$

where $\pi(y)$ is the distribution of distances available for detection and w is a truncation distance, beyond which distances are not considered in the analysis. The above pdf provides the basis of a likelihood from which the parameters of the detection function can be estimated. An important and often overlooked consideration is that $\pi(y)$ is assumed known. This is enforced by design, as the random placement of transects, independently of the animal population, leads to a distribution which is uniform in the case of line transects and triangular in the case of point transects (Buckland et al. 2001).

Given the n distances to detected animals, density can be estimated by

$$\hat{D} = \frac{n\hat{f}(0)}{2L} \quad (2)$$

in the case of line transects with total transect length L, where $f^{(0)}$ is the estimated pdf evaluated at zero distance, and by

$$\hat{D} = \frac{n\hat{h}(0)}{2k\pi} \quad (3)$$

in the case of k point transects, where $h^{(0)}$ is the slope of the estimated pdf evaluated at zero distance (Buckland et al. 2001). This is a useful result because we can then use all the statistical tools that are available to estimate a pdf in order to obtain density estimates. So one can consider plausible candidate models for the detection function and then use standard maximum likelihood to obtain estimates for the corresponding parameters and therefore density estimates.

The most common software to analyze distance sampling data, Distance (Thomas et al. 2010), uses the semi-parametric key+series adjustment formulation from Buckland (1992), in which a number of parametric models are considered as a first approximation and then some expansion series terms are added to improve the fit to the data. Standard model selection tools and goodness-of-fit tests are available for assisting in ▶model selection.

Variance estimates can be obtained using a delta method approximation to combine the individual variances of the random components in the formulas above (i.e., n and either $f^{(0)}$ or $h^{(0)}$; for details on obtaining each component variance, see Buckland et al. 2001). In some of the more complex scenarios, one must use resampling methods based on the non-parametric bootstrap, which are also available in the software.

Given a sufficiently large number of transects randomly allocated independently of the population of interest, estimators are asymptotically unbiased if (1) all animals on the transect are detected, i.e., $g(0) = 1$, (2) sampling is an instantaneous process (typically it is enough if animal movement is slow relative to the observer movement), and (3) distances are measured without error. See Buckland et al. (2001) for discussion of assumptions. Other assumptions, for example that all detections are independent events, are strictly required as the methods are based on maximum likelihood, but the methods are extraordinarily robust to their failure (Buckland 2006). Failure of the $g(0) = 1$ assumption leads to underestimation of density. Violation of the movement and measurement error assumption have similar consequences. Underestimation of distances and undetected responsive movement toward the observers lead to overestimation of density, and overestimation of distances and undetected movement away from the observer lead to underestimation of density. Random movement and random measurement error usually leads to overestimation of density. Naturally the bias depends on the extent to which the assumptions are violated. Most of the current research in the field is aimed at relaxing or avoiding the need for such assumptions. As there are no free lunches in statistics, these come at the expense of more elaborate methods, additional data demands and additional assumptions.

Further details about conventional distance sampling, including dealing with clustered populations, cue counting methods and field methods aspects, can be found in Buckland et al. (2001), while advanced methods, including the use of multiple covariates in the detection function, double platform methods for when $g(0) < 1$, spatial models, automated survey design, and many other specialized topics, are covered in Buckland et al. (2004).

About the Authors

The authors have a wide experience in developing and applying methods for estimation of animal abundance. They have published collectively around one hundred papers in distance sampling methods and applications,

and are actively involved in dissemination of distance sampling methods through international workshops and consultancy. Three of them are also co-authors and editors of the following key books on distance sampling: *Introduction to Distance Sampling: Estimating Abundance of Biological Populations* (Oxford University Press, 2001) and *Advanced Distance Sampling: Estimating Abundance of Biological Populations* (Oxford University Press, 2004).

Cross References

►Statistical Ecology
►Statistical Inference in Ecology

References and Further Reading

Buckland ST (1992) Fitting density functions with polynomials. Appl Stat 41:63–76
Buckland ST (2006) Point transect surveys for songbirds: robust methodologies. The Auk 123:345–357
Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2001) Introduction to distance sampling: estimating abundance of biological populations. Oxford University Press, Oxford
Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2004) Advanced distance sampling. Oxford University Press, Oxford
Thomas L, Buckland ST, Rexstad R, Laake L, Strindberg S, Hedley S, Bishop J, Marques TA, Burnham KP (2010) Distace software: design and analysis of distance sampling surveys for estimating population size. J App Ecol 47:5–14

Distributions of Order *k*

Andreas N. Philippou[1], Demetrios L. Antzoulakos[2]
[1]Professor of Probability and Statistics
University of Patras, Patras, Greece
[2]Associate Professor
University of Piraeus, Piraeus, Greece

The distributions of order k are infinite families of probability distributions indexed by a positive integer k, which reduce to the respective classical probability distributions for $k = 1$, and they have many applications. We presently discuss briefly the geometric, negative binomial, Poisson, logarithmic series and binomial distributions of order k.

Geometric Distribution of Order *k*

Denote by T_k the number of independent Bernoulli trials with success (S) and failure (F) probabilities p and $q = 1 - p$ $(0 < p < 1)$, respectively, until the occurrence of the kth consecutive success. Philippou and Muwafi (1982) observed that a typical element of the event $\{T_k = x\}$ is an arrangement

$$a_1 a_2 \dots a_{x_1+x_2+\cdots+x_k} \underbrace{SS\dots S}_{k}, \tag{1}$$

such that x_1 of the a's are $E_1 = F$, x_2 of the a's are $E_2 = SF, \dots, x_k$ of the a's are $E_k = \underbrace{SS\dots S}_{k-1}F$, and proceeded to obtain the following exact formula for the probability mass function (pmf) of T_k, namely,

$$f(x) = P(T_k = x) = p^x \sum \binom{x_1 + x_2 + \cdots + x_k}{x_1, x_2, \dots, x_k} \left(\frac{q}{p}\right)^{x_1+x_2+\cdots+x_k}, \quad x \geq k, \tag{2}$$

where the summation is taken over all non-negative integers $x_1, x_2, \dots, x_k$ satisfying the condition $x_1 + 2x_2 + \cdots + kx_k = x - k$. Alternative simpler formulas have been derived. The following recurrence for example, due to Philippou and Makri (1985), is very efficient for computations

$$f(x) = f(x-1) - qp^k f(x-1-k), \quad x > 2k \tag{3}$$

with initial conditions $f(k) = p^k$ and $f(x) = qp^k$ for $k < x \leq 2k$. Furthermore, it shows that $f(x)$ attains its maximum p^k for $x = k$, followed by a plateau of height qp^k for $x = k+1, k+2, \dots, 2k$, and decreases monotonically to 0 for $x \geq 2k+1$.

Philippou et al. (1983) employed the transformation $x_i = m_i$ $(1 \leq m_i \leq k)$ and $x = m + \sum_{i=1}^{k}(i-1)m_i$ to show that $\sum_{x=k}^{\infty} f(x) = 1$ (and hence $f(x)$ is a proper pmf). They named the distribution of T_k *geometric distribution of order k with parameter p* and denoted it by $G_k(p)$, since for $k = 1$ it reduces to the classical geometric distribution with pmf $f(x) = q^{x-1}p$ $(x \geq 1)$. It follows from (2), by means of the above transformation and the multinomial theorem, that the probability generating function (pgf) of T_k is given by

$$\phi_k(w) = \sum_{x=k}^{\infty} x^w f(x) = \frac{p^k w^k (1 - pw)}{1 - w + qp^k w^{k+1}}, \quad |w| \leq 1. \tag{4}$$

The mean and variance of T_k readily follow from its pgf and they are given by

$$E(T_k) = \frac{1 - p^k}{qp^k}, \quad Var(T_k) = \frac{1 - (2k+1)qp^k - p^{2k+1}}{(qp^k)^2}. \tag{5}$$

A different derivation of (4) was first given by Feller (1968), who used the method of partial fractions on $\phi_k(w)$ to