THE NEED FOR TEACHING WEIGHTED DISTRIBUTION THEORY: ILLUSTRATED WITH APPLICATIONS IN ENVIRONMENTAL STATISTICS

Lyman L. McDonald
Western EcoSystems Technology, Inc., United States of America
Lmcdonald@west-inc.com

I review weighted distribution theory and applications in observational studies where biased data arise. Recorded observations will be biased and not have the original distribution unless every observation is given an equal chance of being recorded. G. P. Patil and C. R. Rao in 1977 wrote "Although the situations that involve weighted distributions seem to occur frequently in various fields, the underlying concept of weighted distributions as a major stochastic concept does not seem to have been widely recognized." The same quote is applicable today. Our profession is missing an opportunity to provide structure and understanding to a collection of isolated statistical methods for analysis of data from observational studies. Specifically, I present applications in study of wildlife and fish populations that should motivate undergraduate and graduate students of statistics who have an interest in environmental issues.

INTRODUCTION

Often scientists cannot select sampling units in observational studies with equal probability. Well defined sampling frames often do not exist for human, wildlife, insect, plant, or fish populations. Recorded observations on individuals in these populations are biased and will not have the original distribution unless every observation is given an equal chance of being recorded. My first analysis of data from a 'real' study after graduation with a Ph.D. degree in statistics in 1969 involved estimation of the rate at which fish were harvested by sport fisherman during a given time period on a lake in the State of Wyoming, USA. The lake had numerous access points and it was not possible to interview fisherman in a manner that would guarantee equal probability of selection from the finite population of fisherman using the lake during the time period. Interviews had been conducted by a 'roving creel survey' where the interviewer walked around the lake interviewing encountered fisherman and recording how long they had been fishing, how many fish were harvested, and other data. Data collected in this manner are length-biased toward fisherman that fish for longer periods of time. Fisherman who fish for a longer period of time have a higher probability of being interviewed and their observations must be given less weight when estimating population parameters such as the mean number of fish harvested per fishermen. Estimation of other parameters such as the mean length of time a fisherman will fish are related to renewal theory in reliability studies and are more complex. I was ill prepared for the task. Fortunately, I was able to find suitable analysis methods (Robson, 1961) and copy them. Biased data of this type arise in all disciplines of science and scientists and statisticians have discovered and re-discovered ad hoc solutions for correction of the biases.

Weighted distribution theory gives a unified approach for modeling these biased data and should be taught in probability and mathematical statistics courses. Patil and Rao (1977) wrote "Although the situations that involve weighted distributions seem to occur frequently in various fields, the underlying concept of weighted distributions as a major stochastic concept does not seem to have been widely recognized." The same quote is applicable today. Our profession is missing an opportunity to provide structure and understanding to a collection of isolated statistical methods for analysis of data from observational studies.

BASIC NOTATION

Assume that interest is in a random variable X with probability distribution function (pdf) $f_1(x \mid \theta)$, with parameters θ from a given parameter space. Also, assume that the values x and y are observed and recorded in the ratio of $w(x)$:w(y), where $w(x)$ is a non-negative weight function. For example $w(10):w(5) = 2:1$ would imply that $x = 10$ is recorded with twice the probability that $x = 5$ is recorded. If the relative probability that x will be observed and recorded is given by $w(x)$

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

then the probability distribution function of the observed data is $f_2(x \mid \theta) = \frac{w(x) f_1(x \mid \theta)}{c}$, where $c = \sum w(x) f_1(x \mid \theta)$ if *x* is discrete and $c = E_{f_1}[w(x)] = \int w(x) f_1(x \mid \theta) d(x)$, if *x* is continuous. In the 'roving creel survey' mentioned above, *w(x) = x*, where *x* is the length of a fishing trip and we assume the fisherman is contacted at a random point in time. In this case, we say that the observed values of *X* are length- or size-biased and $f_2(x \mid \theta) = \frac{x f_1(x \mid \theta)}{\mu}$, where μ is the mean of *x* over the true, but usually unknown pdf $f_1(x \mid \theta)$.

EXERCISES FOR STUDENTS OF PROBABILITY AND MATHEMATICAL STATISTICS

Basic Exercises

Interesting exercises in probability theory and mathematical statistics can be devised as a means of introducing and teaching weighted distribution theory. For example, assume *X* is Poisson distributed with mean μ and pdf $f_1(x \mid \mu) = \frac{e^{-\mu}\mu^x}{x!}$, $x = 0,1,2,3,...$ If $x = 2$ is observed with twice the probability as $x = 1$, $x = 4$ is observed with twice the probability of $x = 2$, etc., i.e., *w(x) = x* for size-biased sampling, then $f_2(x \mid \mu) = \frac{x}{\mu} \times \frac{e^{-\mu}\mu^x}{x!} = \frac{e^{-\mu}\mu^{x-1}}{(x-1)!}$, $x = 1,2,3,...$ It is impossible to observe $x = 0$, and the entire set of mass points of the Poisson distribution are simply moved over one unit to the right. Given an observed sample of size n from $f_2(x \mid \mu) = \frac{x}{\mu} \times \frac{e^{-\mu}\mu^x}{x!} = \frac{e^{-\mu}\mu^{x-1}}{(x-1)!}$, it will be possible for students to propose an unbiased estimate of μ.

Assume $f_1(x)$ is Poisson with $\mu = 2$, and $f_2(x)$ is Poisson with $\mu = 4$. With discrete random variables and probabilities, it is possible to solve for the weighting function. In this case, $w(x) = cf_2 / f_1 = \frac{c e^{-4} 4^x / x!}{e^{-2} 2^x / x!} = \left(\frac{c}{e^2}\right) 2^x$, the weighting function is proportional to 2^x. To observe a Poisson distribution with mean $\mu = 4$, when sampling from a population with Poisson (mean $\mu = 2$), units with $X=6$ would have to be encountered and recorded with 8 times the probability of encountering units with $X=3$, i.e., in the ratio 64:8 or 8 to 1, etc. These are practical and interesting concepts that are not learned if students are always presented with 'identically and independently distributed random variables.'

Exercises with Continuous Weighted Distributions

Fitness Functions. Continuous weighted distributions are more complex and again present interesting and challenging exercises. One application of continuous weighted distribution theory is in the estimation of 'fitness functions' in the study of natural selection, i.e., estimation of the weighting function, *w(x)*, the relative probability of survival as a function of the characteristic *X* (Manly, 1985). Assume the pdf of *X* for members of a population at time 1 is normal $N(x \mid \mu_1, \sigma_1^2)$ and the pdf of *X* at a later time 2 is normal $N(x \mid \mu_2, \sigma_2^2)$. If both the original and the final observed distributions are normally distributed then the 'fitness function' is given by the exponential function $w(x) = e^{ax+bx^2}$, where $a = \frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}$ and $b = \frac{\left[1/\sigma_2^2\right] - \left[1/\sigma_1^2\right]}{2}$ (O'Donald

1970; Manly, 1985). The interpretation of the 'fitness function' is that members of the population with pdf $N(x \mid \mu_1, \sigma_1^2)$ at time 1 would have to survive with the relative probabilities $w(x) = e^{ax+bx^2}$ in order for the members of the population at time 2 to have the pdf $N(x \mid \mu_2, \sigma_2^2)$. In terms of a sampling experiment, one would have to sample from $N(x \mid \mu_1, \sigma_1^2)$ with the relative probabilities $w(x) = e^{ax+bx^2}$ to generate the observed distribution $N(x \mid \mu_2, \sigma_2^2)$.

Methods for estimation of 'fitness functions' when *X* is a multivariate vector and the distributions $f_1(x)$ and $f_2(x)$ are not specified have been developed since 1985 in the context of estimation of resource selection functions (see below, Manly et al., 2002).

Encounter Sampling. Size-biased data arise in encounter sampling as in the above fisherman example (Otis et al., 1993), the opportunistic interview of shoppers in a super market, encounter of families with genetically transmitted diseases, statistical analysis of failure time data (Kalbfleisch & Prentice, 1980), and many other situations (see, for example, Patil, 2002). Encounter sampling surveys collect data from individuals who are encountered during an activity or behavior (event) by a surveyor that travels a random route through a survey area. When the complete duration time of the event is unavailable, and/or an auxiliary variable is correlated with the duration of the event, traditional estimators taken from sampling theory are inappropriate. Let *X* denote a continuous random variable that represents the duration time of a completed event with pdf f(*x*) and expected value μ_x. The objective is estimation of μ_x. Under the assumption of length-biased sampling, the observed backward or forward duration time of the event will be recorded and denoted *Z*, depending upon whether the starting or ending time of the event is known. Denote the observed pdf of the random variable *Z* by $f_2(z)$. Cox (1969) presented the fundamental assertion that the conditional distribution of either backward or forward observation time *Z* is uniform over the interval [0, *x*] given the complete event duration time $X = x$ and individuals are encountered at a random point in the interval [0,*x*], i.e., $f_1(z) = \frac{1}{x}$. Using this result and standard results from the theory of weighted distributions a good student exercise is to show that $f_2(z) = \frac{1 - F_x(z)}{\mu_x}$, where $F(.)$ is the cumulative distribution function (cdf) corresponding to *f(x)* (Cox, 1969). By substituting the special case z = 0 into this expression we have $F_x(0) = 0$ and Cox's result $\mu_x = 1/f_2(0)$. Thus, an estimator of average duration time of the complete event can be obtained from the empirical pdf of the observed backward or forward observation times, evaluated at the origin. Otis et al. (1993) present estimators for the average value of correlated auxiliary variables measured on encountered individuals.

Distance (Transect) Sampling. An application related to encounter sampling is distance (line transect) sampling for estimation of abundance of wildlife populations (Buckland et al., 2001). In this case, *z* is the observed perpendicular distance to individuals detected with probability *w(z)* (denoted *g(z)* in line transect literature) as an observer travels along a transect line in the study area. Assuming that transect lines are randomly placed in the study area the pdf of detected and non-detected individuals is, $f_1(z) = \frac{1}{W}$, where *W* is the effective half-width of a survey strip, i.e., the maximum possible value of z. The mean probability of detection of an individual in the transect strip of length L and maximum width 2W is $c = \int_0^W w(z) f_1(z) dz = \int_0^W \frac{w(z)dz}{W}$, the normalizing constant in the weighted distribution formula, (probability of detection is assumed to be symmetric on each side of the line). An estimate of c can be obtained by assuming that all units

on the transect line are detected, i.e., w(0) = 1.0, in the relationship $f_2(0) = \frac{w(0)(1/W)}{c}$ or $c = 1/Wf_2(0)$ (Burnham et al. 1980). Thus, an estimator of mean probability of detection of a unit in the transect strip can be obtained from the empirical pdf of the observed distances to detected units, evaluated at the origin. Given that n individuals are detected in the transect strip of length L and width 2W, the Horvitz-Thompson theorem yields an estimate of the abundance of individuals in the strip, $\frac{n}{c} = nWf_2(0)$ (Horvitz and Thompson 1952). Dividing by the area searched, 2WL, the familiar formula for the estimated density of individuals in line transect sampling is obtained, $\frac{nf_2(0)}{2L}$. The literature contains several proposed estimators for $f_2(0)$ and computer packages such as TRANSECT (Burnham et al. 1980) and DISTANCE (Buckland et al. 2001) are available for the necessary computational tasks.

Line transect sampling has been applied to estimate size of many populations of animals including: polar bear (McDonald et al., 1999), birds (Buckland et al., 2001), dolphins (Dawson et al., 2004), brown bear (Quang & Becker, 1996), sea otters, and caribou (Drummer et al., 1990).

Resource Selection Functions. Modeling habitat and food selection by aquatic and terrestrial animals using weighted distribution theory and Resource Selection Functions (RSF) were first introduced by McDonald et al. (1990). Given a vector, *X*, of variables measured on sampling units, the RSF is denoted by the weighting function *w(x)* in the weighted distribution $f_2(x) = w(x)f_1(x)/c$. The function $f_1(x)$ denotes the pdf of *X* for resource units available to the animals and $f_2(x)$ denotes the pdf of *X* for resource units selected by the animals. Thus, *w(x)*, is a function proportional to the probability of a unit with $X=x$ being selected by the animal(s) from the available population of units in a study area.

Given a sample from the population of units in a study area, i.e. from $f_1(x)$, and a separate sample from the population of units selected by the animal(s), i.e. from $f_2(x)$, then sufficient information exists to estimate the third member, the RSF *w(x)*, of the relationship. However, as usual, the devil is in the details, particularly when estimating the ratio of pdfs for continuous variables, $w(x) = \frac{cf_2(x)}{f_1(x)}$. Various methods have been devised for fitting models to the RSF (Manly et al., 2002; Johnson et al., 2006; Thomas & Taylor, 2006). Probably the most common method involves fitting an exponential model, $w(x) = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$, the correct model if both distributions happen to be approximately normally distributed. The primary advantage of a RSF is that clear statements can be made concerning the relative probability that each of several food or habitat units will be selected. Given a model for *w(x)*, maps of a study area can be prepared showing which units were selected with higher probability and presumably are most valuable for survival or reproduction of the animals. Obviously, the modeling can be complex depending on the age and sex of study animals, if they are solitary or not, etc.

There is active mathematical/statistical research being conducted on estimation of RSFs using the weighted distribution approach with continuous random variables. If the normalizing constant $c = E_{f_1}[w(x)] = \int w(x)f_1(x)d(x)$ can be estimated then *w(x)* can be standardized so that $w^*(x) = w(x)/c$ is a Resource Selection "Probability" Function. Technically, it is feasible to do so now, with the availability of Geographical Information Systems (GIS) that can provide data on very large samples of units in a study area and store-on-board Global Positioning System (GPS) radiocollars attached to animals and programmed to obtain a latitude-longitude location every few minutes. Lele (2008) and Johnson et al. (2008) independently suggest stepwise procedures for estimation of c. They consider a model for *w(x)*, e.g., an exponential function, and fit the model using a very large sample of units from the study area and a large sample of units used by an animal. Next, average *w(x)* over the sample of units from the

study area to obtain an estimate of c, say $\hat{c}$. Refit $w(x)/\hat{c}$, and repeat the process until the model converges to stable values. Interestingly, the interpretation of estimates of relative probability of selection of discrete study units provided by an RSF are easy, however interpretation of "probability" statements based on $w(x)/\hat{c}$ are elusive.

CONCLUSION

Many observational studies lack a well defined sampling frame for selection of sampling units, particularly in observational studies of humans, wildlife, insects, fisheries, and plants. The studies share the characteristic that it is often not possible to make a list of the units of interest and select a sample with equal or known probabilities. Weighted distribution theory provides a unifying approach for correction of biases that exist in unequally weighted sample data. Also, the theory provides a means of fitting models to the unknown weighting function when samples can be taken both from the original distribution and the resulting 'biased' distribution. These problems exist in all disciplines of science and numerous ad hoc solutions have been developed.

Unfortunately, authors of textbooks for mathematical statistics have largely ignored weighted distribution theory. I believe this is partially due to the fact that 'non-identically distributed' data do not fit nicely into the classical system of mathematical statistics. Teachers of statistics have the opportunity to include weighted distribution theory in text books, to present elementary exercises and interesting applications, and to better prepare their students to deal with observational studies.

Also, unfortunately, there is no good source for material that is ready made for the classroom. Good starting points from which to develop course material are the works of Professor G. P. Patil, Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A. (e.g., Patil 2002; Patil & Rao, 1977, 1978).

REFERENCES

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. (2001). *Introduction to Distance Sampling.* Oxford: Oxford University Press.

Burnham, K. P., Anderson, D. R., & Laake, J. L. (1980) Estimation of density from line transect sampling of biological populations. *Wildlife Monographs, 72*, 1-202.

Cox, D. R. (1969). Some sampling problems in technology. In N. L. Johnson & H. Smith (Eds.), *New developments in survey sampling* (pp. 506-527). New York: John Wiley & Sons.

Dawson, S., Slooten, E., DuFresne, S., Wade, P., & Clement, D. (2004). Small-boat surveys for coastal dolphins: line-transect surveys for Hector's dolphins (Cephalorhynchus hectori). *Fishery Bulletin, 102*, 441-451.

Drummer, T. *D.*, Degange, A. R., Pank, L. L., & McDonald, L. L. (1990). Adjusting for Group Size Influence in Line Transect Sampling. *The Journal of Wildlife Management, 54*, 511-514.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association.* 47: 663- 685.

Johnson, C. J., Nielson, S. E., Merrill, E. H., McDonald, T. L., & Boyce, M. S. (2006). Resource selection functions based on use-availability data: Theoretical motivation and evaluation methods. *Journal of Wildlife Management, 70,* 347-357.

Johnson, D. S., Thomas, D. L., Ver Hoef, J. M., & Christ, A. (2008). A general framework for he analysis of animal resource selection from telemetry data. *Biometrics, 64*, 968-976.

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The Statistical analysis of failure time data.* New York: John Wiley & Sons.

Lele, S. R. (2008) A new method of estimation of resource selection probability function. *Journal of Wildlife Management, 73*, 122-127.

McDonald, L. L., Manly, B. F. J., & Raley, C. M. (1990) Analyzing foraging and habitat use through selection functions. *Studies in Avian Biology, 13,* 325–331.

McDonald, L. L., Garner, G. W., & Robertson, D. G. (1999) Comparison of aerial survey procedures for estimating polar bear density: Results of pilot studies in Northern Alaska. In G. W. Garner, S. C. Amstrup, J. L. Laake, B. F. J. Manly, L. L. McDonald & D. G. Robertson

(Eds.), *Marine Mammal Survey and Assessment Method* (pp. 37-51). Rotterdam: A. A. Balkema.

Manly, B. F. J. (1985). *The Statistics of Natural Selection on Animal Populations.* London: Chapman and Hall.

Manly, B. F. J., McDonald, L. L., Thomas, D. L., McDonald, T. L., & Erickson, W. P. (2002). *Resource Selection by Animals: Statistical Design and Analysis for Field Studies, Second Edition*, Dordrecht: Kluwer Academic Publishers.

O'Donald, P. (1970). Change of fitness by selection for a quantitative character. *Theoretical Population Biology, 1*, 219-232.

Otis, D. L., McDonald, L. L., & Evans, M. (1993). Parameter estimation in encounter sampling surveys. *Journal Wildlife Management, 57*, 543-548.

Patil, G. P. (2002). Weighted Distributions. In A. H. El-Shaarawi, & W. W. Piegorsch (Eds.), *Encyclopedia of Environmetrics* Volume 4 (pp. 2369-2377). Chichester: John Wiley & Sons.

Patil, G. P., & Rao, C. R. (1977). Weighted distributions: a survey of their application. In P. R. Krishnaiah (Ed.), *Applications of Statistics* (pp. 383-405). North Holland Publishing Company.

Patil, G. P., & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics, 34*,179-180.

Thomas, D. L., & Taylor, E. J. (2006). Study designs and tests for comparing resource use and availability II. *Journal of Wildlife Management, 70*, 324–336.

Quang, P. X., & Becker, E. F. (1996). Line Transect Sampling Under Varying Conditions with Application to Aerial Surveys. *Ecology, 77*, 1297-1302.

Robson, D. S. (1961). On the statistical theory of a roving creel census of fishermen. *Biometrics, 17*, 415-437.

I review weighted distribution theory and applications in observational studies where biased data arise. Recorded observations will be biased and not have the original distribution unless every observation is given an equal chance of being recorded. G. P. Patil and C. R. Rao in 1977 wrote â€œAlthough the situations that involve weighted distributions seem to occur frequently in various fields, the underlying concept of weighted distributions as a major stochastic concept does not seem to have been widely recognized.â€ The same quote is applicable today. Our profession is missing an opportunity to Environment statistics is the application of statistical methods to environmental science. It covers procedures for dealing with questions concerning the natural environment in its undisturbed state, the interaction of humanity with the environment, and urban environments. The field of environmental statistics has seen rapid growth in the past few decades as a response to increasing concern over the environment in the public, organizational, and governmental sectors. Extreme value theory provides the solid fundamentals needed for the statistical modelling of such events and the computa-tion of extreme risk measures. The focus of the paper is on the use of extreme value theory to compute tail risk measures and the related condence intervals, applying it to several major stock market indices.Â Figure 4 illustrates the shape of the generalized Pareto distribution GÎ¾,Ïƒ(x) when Î¾, called the shape parameter or tail index, takes a negative, a positive. and a zero value. The scaling parameter Ïƒ is kept equal to one.Â 4 Application. Our aim is to illustrate the tail distribution estimation of a set of nancial series of daily returns and use the results to quantify the market risk. Table 1 gives the list of the nancial series considered in our analysis.