Trim: 247mm × 174mm Top: 12.653mm Gutter: 16.871mm
CUUK1798-14 CUUK1798/Gitzen *et al.* ISBN: 978 0 521 19154 8 January 6, 2012 12:58

14 Survey analysis in natural resource monitoring programs with a focus on cumulative distribution functions

Thomas M. Kincaid and Anthony R. Olsen

Introduction

Typical objectives for environmental resource monitoring programs include estimating the current status of the resource, estimating change in status between two time periods, and estimating trends in status over time. For most monitoring programs, status is estimated using multiple indicators determined for each sample site that are derived from measurements of biological, chemical, and physical attributes obtained at the site. What is meant by estimating current status? First, the estimate applies to a specified portion of the region included in the monitoring program, typically the entire region. Second, a specific summary measure must be chosen. Common summary measures are estimates of the population mean, the percentage of the population that is less than a particular value (e.g. percentage of water bodies meeting a water quality standard or percentage of a dry forest region that has tree densities $< x$ trees/ha), the population median, the percentage of the population occurring in categories, or the population standard deviation for an indicator (measured attribute).

An additional summary measure is an estimate of the population cumulative distribution for the indicator. The population cumulative distribution is simply the percentage of the population that is less than or equal to each possible value of an indicator. The population cumulative distribution provides complete information about the indicator distribution (Box 14.1). It has the advantage that it can be used for both categorical and continuous data. It is common in statistics to call the cumulative distribution the cumulative distribution function (CDF) and we will adopt that convention in this chapter. Measurements for indicators obtained from a probability survey can be used to estimate these summary measures of status. For complex survey designs that employ stratification or unequal probability of selection, estimation of current status for any of these summary measures must use weights that are derived from the stratification or unequal probability of selection used in the design.

The material presented in this chapter is applicable to natural resources that can be classified as either finite (discrete) or continuous. A finite resource consists of a collection of distinct units, such as lakes or stream reaches, within a study region or

Design and Analysis of Long-term Ecological Monitoring Studies, ed. R.A. Gitzen, J.J. Millspaugh, A.B. Cooper and D.S. Licht. Published by Cambridge University Press. © Cambridge University Press 2012.

Box 14.1 Take-home messages for program managers

Implementing an environmental resource monitoring program based on a spatial survey design not only includes choosing an appropriate survey design to meet a program's monitoring objectives but also includes ensuring that the statistical analyses are appropriate for the survey design. This chapter focuses on the utility of the population cumulative distribution as a summary measure for a monitoring variable. The population cumulative distribution is simply the percentage of the population that is less than or equal to each of all possible values of a variable. As a measure of resource status, the population cumulative distribution provides complete information about the distribution of values of a variable, which can be especially useful in management contexts. A natural-resource agency may have a management-threshold goal, e.g. percent of the stream miles with water temperature below 22°C. While the threshold goal is 22°C, how the percentage changes if the threshold was different is valuable to know as well. The cumulative distribution provides this information. Estimating the cumulative distribution based on sample data requires knowing the properties of the survey design, e.g. if it was stratified or used unequal probability of selection. These more complex survey designs require a more complex statistical analysis. For example, with more complex designs it is no longer appropriate to simply tabulate the number of sites with water temperature below 22°C and divide by the number of sites to estimate the proportion of stream miles below 22°C. A weight for each site, based on the survey design, must be used to calculate a weighted mean. Software (a free R package) is available for appropriate estimation of CDFs and comparisons of CDFs (e.g. at two time points or for two different subregions).

spatial domain. An example of a finite resource is the set of lakes in California, USA, with a surface area > 1 ha. Each lake will have a single value for an indicator such a fish assemblage index of diversity. An example of a continuous resource is the collection of perennial streams in Oregon. Conceptually; the value of an indicator varies continuously throughout the stream linear network, e.g. pH of the water or density of fish per 100 m. To define the continuous resource population in this example, streams are modeled as having no width. Another example of a continuous resource is the land area within a national forest. An indicator of interest may be an index of vegetation diversity which would vary continuously throughout the national forest. Another aquatic example of a continuous resource is the collection of estuaries for the State of Oregon. In this case it is assumed that the indicators (e.g. dissolved oxygen concentration) vary not only across the estuaries but also within an estuary. The latter two examples are examples of an areal continuous resource. In geographic information systems (GIS) terminology, a finite resource is a point GIS layer, a linear network is a linear layer and an areal resource is a polygon layer (see also Chapter 6). For all types of environmental resources, we will use the term “extent” to reference the size of the resource, i.e. the number of discrete units for a finite resource, total length for a linear resource, and total area for an areal resource.

This chapter is divided into five sections. The first section introduces population summary measures and procedures for estimation with a focus on the CDF. The second section focuses on using the sample CDF to estimate population percentiles. The third section introduces deconvolution as a way to remove the effect that measurement error associated with indicators has on CDF estimates. The fourth section introduces hypothesis testing for comparing two CDFs. The fifth section briefly discusses unresolved issues in need of further research or development.

Population summary measures and their estimation

The Horvitz–Thompson theorem (Horvitz and Thompson 1952) and its generalization is the statistical basis for estimation of population summary measures. For a function z defined on a finite resource $\boldsymbol{U}$, the theorem provides a procedure for estimating the population total for z,

$$T_Z = \sum_{x \in \boldsymbol{U}} z(x). \tag{14.1}$$

Note that the term “population”, as used in this chapter, references the set of values of z (either the indicator value directly or a function of the indicator) defined on the resource $\boldsymbol{U}$ (the target universe). An example may make this more concrete. Let the universe be the set of all lakes in Oregon and x be the indicator lake pH, and assume the function z identifies lakes with pH less than 5 ($z = 1$ if pH <5 and $z = 0$ otherwise). The population total is then simply the number of lakes with pH < 5. If the total number of lakes in the population is known, then the percentage of lakes with pH < 5 is simply the population total divided by the total number of lakes. If the function z is simply equal to x, then an estimate for the mean pH is the population total divided by the total number of lakes.

For a sample S selected from $\boldsymbol{U}$ such that, for every element $x \in \boldsymbol{U}$, there is a positive probability, $\pi(x)$, that the element x is included in S, an estimate of the total is

$$\hat{T}_Z = \sum_{x \in S} \frac{z(x)}{\pi(x)}. \tag{14.2}$$

The quantity $\pi(x)$ is the inclusion probability for an element, and a sample S that is selected with a positive inclusion probability for every element is called a probability sample (see also Chapters 5 and 6). The inverse of the inclusion probability is called the survey design weight. It can be seen that $\hat{T}_Z$ is a weighted sum of the sample values of the function z, where weights are the survey design weights. Horvitz and Thompson (1952) also provided a formula for the variance of $\hat{T}_Z$. The variance formula requires the pairwise inclusion probability, $\pi(x_i, x_j)$, which is the probability that both elements x_i and x_j are included in the sample S. (Note that Chapter 5 provides a function for estimating these pairwise inclusion probabilities.) Under the condition that $\pi(x_i, x_j)$ is positive for every pair of elements x_i and x_j in $\boldsymbol{U}$, Horvitz and Thompson (1952) provided a formula for an unbiased estimator of the variance of $\hat{T}_Z$.

Cordy (1993) developed an extension of the Horvitz–Thompson theorem to sampling from a continuous resource. For this case, the population total is given by

$$T_Z = \int\limits_{\boldsymbol{U}} z(x)dx, \tag{14.3}$$

where the summation used previously is replaced by integration. Cordy (1993) defined an inclusion density function for a continuous resource that is analogous to the inclusion probability for a finite resource. He also established that the estimator $\hat{T}_Z$ is applicable when sampling from a continuous resource with $\pi(x)$ referring to the inclusion density instead of the inclusion probability.

For a target universe corresponding to an environmental resource $\boldsymbol{U}$, the cumulative distribution function, CDF, for a random variable z defined on $\boldsymbol{U}$ is given by $F_Z(t) = \Pr(z \leq t)$ for all t. For a probability sample S, an estimate of the CDF is given by

$$\hat{F}_Z(t) = \frac{\sum\limits_{x \in S} \frac{I(z(x) \leq t)}{\pi(x)}}{\sum\limits_{x \in S} \frac{1}{\pi(x)}}, \tag{14.4}$$

where the indicator function, I, is defined as

$$I(z(x) \leq t) = \begin{cases} 1, & z(x) \leq t \\ 0, & \text{otherwise} \end{cases}.$$

The expressions in both the numerator and denominator of $\hat{F}_Z(t)$, Equation (14.4), are Horvitz–Thompson estimators. The numerator expression estimates the extent (i.e. size) of the spatial domain for which values of the function z are less than or equal to the value t. The denominator expression, which is the sum of the survey design weights for the sample, estimates the entire resource extent (i.e. size). If an estimate of the mean of an indicator is desired, then the estimate is given by

$$\hat{\hat{x}} = \frac{\sum\limits_{x \in S} \frac{x}{\pi(x)}}{\sum\limits_{x \in S} \frac{1}{\pi(x)}} \tag{14.5}$$

where the numerator is the weighted sum of the indicator values and the denominator is the sum of the weights. Diaz-Ramos *et al*. (1996) gives detailed examples on the estimation of CDFs and other summary measures of status.

As an introduction to CDF estimation, consider the CDFs displayed in Fig. 14.1. The lakes used to create the figure were extracted from the sample frame that was employed for the National Lakes Assessment (USEPA 2009b) in the US. Specifically, the 818 lakes that were included in the National Eutrophication Survey, NES (USEPA 1975), were retained. This set of lakes will be used as the resource of interest. An index of biotic integrity (IBI) for lakes was used as the response variable z for constructing the CDFs. For illustration purposes, use of data with known properties is helpful; therefore, we simulated IBI data for all lakes from a Normal distribution with mean 50 and standard

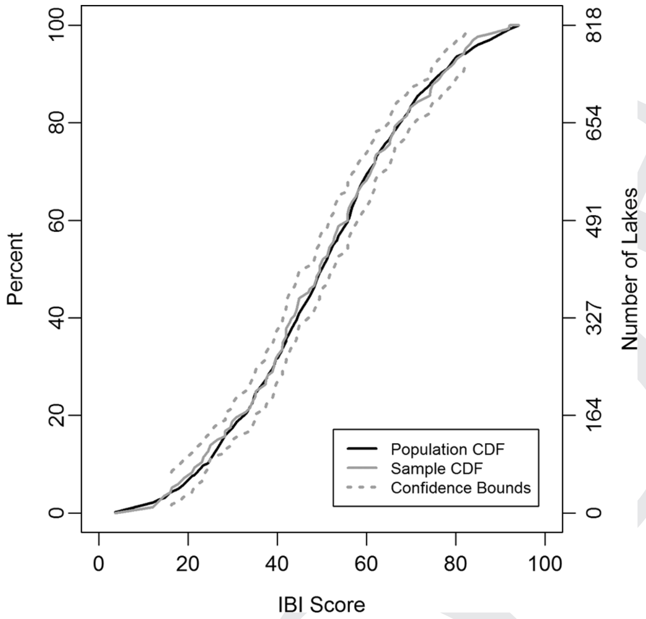

Figure 14.1 Simulated index of biotic integrity (IBI) CDFs for lakes in the US National Eutrophication Survey that were included in the survey frame for the National Lakes Assessment. Shown are the known (simulated) CDF for all of the lakes in the survey (the population CDF) and the CDF and its associated confidence bounds estimated from a sample of lakes (identified as the sample CDF).

deviation 20. The IBI CDF for all of the NES lakes (i.e. the population CDF) is therefore known in this case. A sample of lakes was selected from the population, and the IBI CDF and associated confidence bounds were estimated (Fig. 14.1). Because they tend not to be very informative, confidence bounds for the tails of the distribution are not displayed. Note that the confidence bounds enclose the population CDF.

Estimation of the CDF requires knowledge of the survey design used to select the sample of lakes and matching the survey analysis to that design, which can be challenging in many monitoring situations (Box 14.2). For our example, we used an unequal probability survey design where the unequal probability of selection was based on 4 lake area size classes with 25 lakes selected in each for a total sample size of 100 lakes. We implemented spatially balanced sampling using the generalized random tessellation stratified (GRTS) design (Stevens and Olsen 2004; see Chapter 6). We implemented the design using the *grts* function in the “spsurvey” package for R (R Development Core Team 2010) demonstrated in Chapter 6, although other options are available. The “spsurvey” package is available on the R CRAN website (http://www.r-project.org/) and the US Environmental Protection Agency’s Aquatic Resource Monitoring (ARM) website (http://www.epa.gov/nheerl/arm/).

The IBI values for the 100 selected lakes were then used to estimate the population CDF using the *cdf.est*() function in the “spsurvey” package:

Box 14.2 Common challenges: design and analysis issues

Estimating summary measures, such as the population CDF, and their standard errors or confidence intervals is computationally easily accomplished with available software. Small sample sizes, especially associated with a complex design, can result in large standard errors and consequently wide confidence intervals. If the weights associated with the design are positively correlated with the values of the indicator, i.e. sites with large weights have large indicator values, then the standard errors will be small. For example, this would occur if stratification was used to separate the population into homogeneous strata with the goal of reducing the variance. Estimating the population CDF can also be impacted by 1 or 2 observations that are extreme and also have large weights. A general rule of thumb is to have the weights vary by less than a factor of 100, which can be difficult to accomplish in some cases. Increasing the sample size and reducing the range of the weights will result in lower standard errors and shorter confidence intervals.

Monitoring programs may implement a monitoring design where it takes multiple years (5–10) before accumulating enough sites for any meaningful CDF or other summary measure estimate can be made. Accumulating enough sites can be accomplished by using any number of panel designs (see Chapters 7, 8, and 10). Estimating the CDF requires specifying a temporal period, e.g. 2005–2009, that defines the population in time. Having data from multiple years adds another source of variation to all the sources of variation included in measurement error, potentially increasing bias in CDF estimates. Some panel designs include revisits to the same sites within the temporal period. When this is the case, a question arises on how the data from these multiple visits is used in the estimation process. One approach is to average the values and then use the average in the estimation of the CDF. This results in the sites with averages having smaller measurement errors than sites that have no revisits. Consequently, the CDF estimate is impacted since the convolution of the measurement error with the distribution of interest differs for values that are averaged and those that are not. An alternate solution is to use only one of the revisit values and ignore the remaining revisits. More complex analyses based on linear models could be used to include all the data without averaging.

```
CDF_Sample <- cdf.est(z=NESsample$IBI,
      wgt=NESsample$wgt,
      x=NESsample$xcoord, y=NESsample$ycoord)
```

where “z” is the IBI data, “wgt” is the survey design weights for the lakes, and “x” and “y” are the lake location coordinates. The *cdf.est*() function implements the Horvitz–Thompson estimator $\hat{F}_Z(t)$ to calculate the CDF estimate. This estimator produces values that range between 0 and 1. Typically, the estimate is multiplied by 100 to express the estimate as a percentage. Since *cdf.est* by default uses the local mean variance estimator (Stevens and Olsen 2003; see also Chapter 7) to calculate standard error (se) of the CDF estimate, it is necessary to provide the lake location information. An option also exists to use the standard Horvitz–Thompson variance estimator. Confidence bounds, $\hat{F}_Z(t) \pm z^* se$, are calculated for each unique value in the sample and connected pointwise to display the confidence bounds (Fig. 14.1).

The population CDF can also be estimated in terms of the resource extent, i.e. in this example in terms of the number of lakes in the population with indicator values $\leq t$. For a linear or areal resource the estimate would be in terms of length or area instead of number of finite units. Two options exist when estimating the population CDF in terms of the resource extent. First, if the resource extent is known, then the CDF estimate is calculated using a ratio estimator: the product of the known resource extent and the ratio of Horvitz–Thompson estimators [Equation (14.4)] used to calculate the CDF estimate for proportion of a resource. This first CDF estimate in terms of resource extent will range between zero and the resource extent. If the resource extent is not known, the CDF estimate in terms of the resource extent is calculated using the Horvitz–Thompson estimator in the numerator of Equation (14.4). This second type of CDF estimate for total of a resource will range between zero and the sum of the survey design weights, which is an estimate of the resource extent. The *y*-axis displayed on the right side of Fig. 14.1 expresses the CDF as the resource extent. Note that the maximum value is 818 lakes, the known resource extent.

Percentile estimation

The (100*p*)th percentile for a random variable *z* defined on an environmental resource ***U*** is given by the smallest value ξ_p such that $F_Z(\xi_p) \geq p$. Note that ξ_p will belong to the range of values allowed for the response variable *z*. Percentile estimates are calculated by inverting the estimated sample CDF, $\hat{F}_Z(t)$. Specifically, the percentile estimate, $\hat{\xi}_p$, is provided by the value *t* such that $\hat{F}_Z(t) = p$. Note that *t* must be one of the response values at which the CDF was estimated. When no value *t* is available that satisfies the equality, linear interpolation is used to calculate $\hat{\xi}_p$. Confidence bounds for $\hat{\xi}_p$ are calculated by inverting the confidence bounds for $\hat{F}_Z(t)$. Note that the upper bound for $\hat{F}_Z(t)$ provides the lower bound for $\hat{\xi}_p$ and conversely regarding the lower bound of $\hat{F}_Z(t)$. As a result of the procedure used to calculate confidence bounds, standard error estimates for $\hat{\xi}_p$ are not calculated.

Measurement error and CDF deconvolution

We want the estimate for the population CDF will be an unbiased estimate of the true population CDF. The true population CDF reflects the frequency distribution of an indicator of interest for an environmental resource when the indicator is measured without error. This frequency distribution reflects the variation of the indicator across the resource and is the distribution that is of interest. For convenience, we refer to this variation as site-to-site variation. The survey design, measurement process, and components of natural variation introduce additional variation that can lead to biased estimation of the CDF (see Chapters 7 and 8). Overton (1989) classified the additional variance as measurement errors. Kincaid *et al*. (2004) used the term "extraneous variance" to reference these "measurement errors". We adopt the term "measurement error" where it includes all the sources of variation involved in obtaining a value for an indicator at a single sample site, comparable to the use of this term in previous chapters. Although our interest is in the distribution of an indicator "x", we actually observe "y" where $y = x + \varepsilon$ and ε is the measurement error. The observed indicator values convolute the measurement error variation with the true population variation resulting in the observed distribution having greater variation than the underlying true population. This produces a bias in the CDF estimate, which means that the CDF estimate will have a greater proportion of the estimate in the tails of the CDF compared to an unbiased CDF estimate.

Urquhart *et al*. (1998) and Larsen *et al*. (2001) described a variance model that addresses major sources of variation affecting status estimation and trend detection in monitoring surveys (see Chapters 7–10). This model identified four variance components: (i) site variance, (ii) year variance, (iii) site-by-year interaction variance, and (iv) residual variance. Site variance is site-to-site (spatial) variation in the population. Year variance is the year-to-year variation affecting all sites equally. Site-by-year interaction variance is year-to-year variation affecting individual sites that is not accounted for by year variance. Residual variance is the remaining variation not accounted for by the other variance components. See Larsen *et al*. (1995) for discussion regarding the components of residual variance, i.e. measurement variance.

The relative magnitude of measurement variance to the true population variance determines the amount of bias in the CDF estimate. The relationship between the amount of measurement variance and bias in the CDF was considered by Overton (1989). Bias resulting from measurement variance is illustrated in Fig. 14.2 using a simulated species richness variable. The upper portion of Fig. 14.2 displays CDF estimates, and the lower portion displays corresponding probability density estimates. Densities are included since it is easier to see that the effect of measurement variance is to increase the variation in the observed distribution. Density estimates were calculated using an R function that implements the average shifted histogram (ASH) algorithm (Scott 1985). Data for the species richness variable was simulated using the Normal distribution with mean 20 and standard deviation 5. The other variables displayed in the figure were created by adding 25%, 50%, and 100% measurement variance to the original variable. As an example, to create the variable with 100% measurement variance, data simulated using the Normal distribution with mean 0 and standard deviation 5 was added to the species simulated richness variable.

As can be observed, measurement error causes the CDF and density estimates to occur across a greater range of species richness values in comparison to the CDF estimate with no measurement error. In addition, bias in the CDF and density estimates increases as measurement variance increases. Note also that bias in the CDF is not constant but is greatest approximately midway between the median value (50th percentile) and the tails of the distribution. For species richness equal to 15, the CDF estimate for the original variable is 11.1%, which increases to 14.2%, 15.9%, and 20.4% for 25%, 50%, and 100% measurement variance, respectively.

Convolution is a term that refers to a variable that is a mixture of two or more distributions. In the context of this chapter, a variable that contains measurement error

Trim: 247mm × 174mm Top: 12.653mm Gutter: 16.871mm
CUUK1798-14 CUUK1798/Gitzen *et al.* ISBN: 978 0 521 19154 8 January 6, 2012 12:58

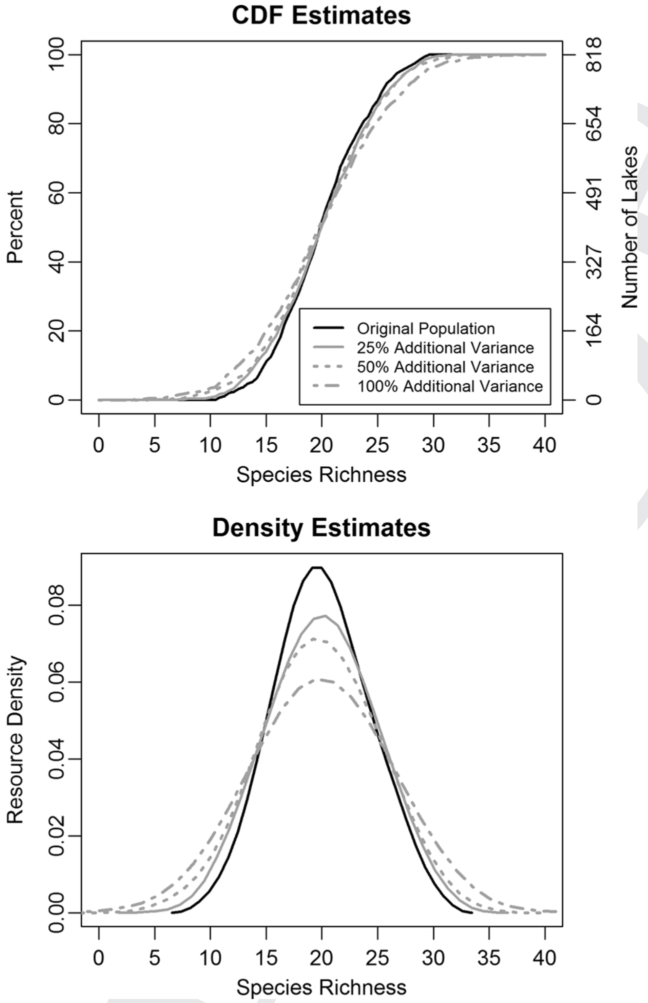

Figure 14.2 Illustration of the effect of extraneous variance on the simulated population CDF and population probability density function for IBI score for lakes in the National Eutrophication Survey that were included in the survey frame for the National Lakes Assessment.

is the convolution of the distribution of the variable of interest and a distribution that reflects the measurement error. Deconvolution is the name for a process that removes from the CDF estimate the bias caused by measurement error. Although deconvolution will remove bias from the CDF estimate, the cost associated with deconvolution is increased width of confidence bounds for the deconvoluted CDF estimate. Note that the removal of bias and the increased confidence interval width result in confidence intervals that better reflect their stated confidence level. Kincaid *et al*. (2004) discussed the cost associated with deconvolution and assessed the cost using two measures: (i) increase in the width of CDF confidence bounds, and (ii) increase in the sample size required to achieve confidence bounds equivalent to presence of no measurement error.

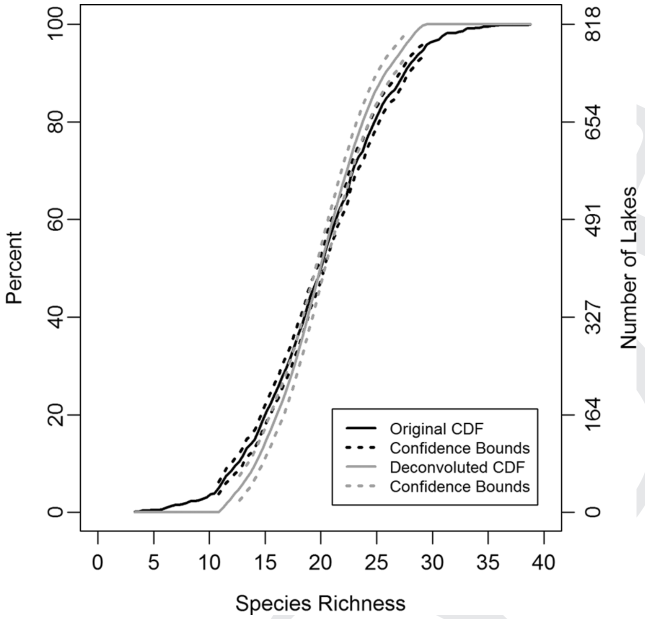

Figure 14.3 Illustration of deconvolution using simulated species richness CDFs for lakes in the National Eutrophication Survey that were included in the survey frame for the National Lakes Assessment. Shown are the CDF for species richness with extraneous variance (identified as the original CDF) and its associated confidence bounds, and the deconvoluted CDF and its associated confidence bounds (identified as the deconvoluted CDF).

Estimated CDFs using deconvolution are illustrated in Fig. 14.3. The simulated species richness variable to which 100% measurement variance was added will be used as the response variable. The *cdf.est*() function in "spsurvey" was used to calculate the CDF for the original variable with added variance. The *CDF.decon*() function in "spsurvey" was then used to calculate the deconvoluted CDF. The deconvolution process implemented in the *CDF.decon* procedure is based on Stefanski and Bay (1996). As discussed previously, tails of the deconvoluted CDF are shifted towards the distribution center. Increased width of confidence bounds for the deconvoluted CDF in comparison to the original CDF can be observed in Fig. 14.3.

Hypothesis testing for CDFs

Although extensive methodology exists for inference about CDFs in the context of simple random sampling, there is relatively little literature that addresses inference about the CDF for sampling designs involving stratification and unequal probability of selection or more complex designs. For simple random sampling the Kolmogorov–Smirnov (KS) and Cramér–von Mises (CvM) statistics typically are used for inference about CDFs. Use of the KS and CvM statistics for testing CDFs is discussed in numerous books, e.g. Conover (1980).

Trim: 247mm × 174mm Top: 12.653mm Gutter: 16.871mm
CUUK1798-14 CUUK1798/Gitzen *et al.* ISBN: 978 0 521 19154 8 January 6, 2012 12:58

As a means for comparing CDFs from complex sample surveys, Krieger and Pfeffermann (1997) use an approach based on grouping the data into a fixed number of intervals and using the standard Pearson chi-squared test for categorical data analysis (χ^2) to conduct inference. The χ^2 test assumes multinomial sampling, i.e. simple random sampling. When applied to a complex survey design, χ^2 does not perform adequately. For complex survey designs, the Wald (1943) statistic provides a mechanism for incorporating features of the design into categorical data analysis. Rao and Scott (1981) and Scott and Rao (1981) examined the impact of stratification and clustering on performance of χ^2 for testing goodness of fit and independence in two-way contingency tables and developed corrections to the Pearson statistic. Rao and Scott (1981) developed first-order and second-order corrections to χ^2. Rao and Thomas (1988, 1989) present detailed descriptions of the Wald and Rao–Scott test statistics. The Wald and Rao–Scott test statistics are referenced to the chi-squared distribution. Thomas *et al.* (1996) developed modifications to the Wald and Rao–Scott tests that are referenced to the F distribution.

Kincaid (2000) conducted simulations on the performance of the Wald and Rao–Scott test statistics to hypothesis testing for statistical difference between CDFs. To utilize the tests, the CDFs were classified into a set of non-overlapping classes, and the statistics were calculated. Simulation results presented in Kincaid (2000) indicated a strong tendency for statistical power to increase as the number of classes was decreased. In subsequent simulations (T. Kincaid, unpublished data) the F distribution versions of the statistics show greater power in comparison to versions using the chi-squared distribution. Overall, the conclusion is that the F-based version of the Wald statistic is the best choice to use for inference regarding CDFs.

The *cdf.test*() function in the “spsurvey” package for R calculates the chi-squared distribution and F distribution versions of the Wald and Rao–Scott test statistics. The *cdf.test*() function was used to test for differences among IBI score CDFs for subpopulations defined by the lake size classes used for the survey design described previously. In order to provide an illustration of the test, random noise from the Uniform distribution was added to the IBI scores for the smallest lake size class. The noise served to shift the distribution for the smallest lakes toward larger values. As expected (Table 14.1), the smallest lake size class was significantly different from the other lake size classes for the standard CDFs. None of the other lake size classes were significantly different.

Future research and development

Although procedures for estimating CDFs are well developed, some aspects of estimating confidence intervals and testing of hypotheses need additional work. Estimating confidence limits at either end of the cumulative distribution, e.g. say less than 5% or greater than 95%, by relying on $\hat{F}_Z(t) \pm z^{*}se$ is known to be problematic in giving confidence intervals with the expected confidence level. A general solution to the problem is not currently available.

Although the procedures for testing for differences between CDFs that were discussed provide a mechanism for accommodating complex survey designs, discretizing the CDFs

Table 14.1 Results of example hypothesis test for differences between simulated IBI CDFs for subpopulations defined by lake size classes for lakes in the National Eutrophication Survey that were included in the survey frame for the National Lakes Assessment.

Lake size class 1 (ha)	Lake size class 2 (ha)	Wald statistic	*P*-value
0–200	200–1000	4.570	0.002
0–200	1000–5000	3.120	0.027
0–200	> 5000	3.463	0.018
200–1000	1000–5000	0.517	0.566
200–1000	> 5000	0.871	0.401
1000–5000	> 5000	0.005	0.995

by dividing them into a set of classes is less than ideal. Developing test statistics that treat the CDF as a whole is a future research objective.

Summary

This chapter focuses on the utility of the population cumulative distribution as a summary measure for variables measured by a long-term monitoring program. The population cumulative distribution provides complete information about the variable distribution. Estimating the cumulative distribution based on sample data requires knowing the properties of the survey design, e.g. if it was stratified or used unequal probability of selection. These more complex survey designs require a more complex statistical analysis.

In this chapter we define the population cumulative distribution, provide estimators of the population cumulative distribution and associated standard errors and confidence intervals, and introduce tests to determine if the cumulative distributions from two time periods or two subregions are different. While calculating the estimates and tests are more complex, open source software is available to complete the calculations. The chapter also discusses the impact (potential bias) of measurement error on the estimated cumulative distribution and presents a solution (deconvolution) to remove the bias.

Acknowledgments

The information in this document has been funded by the US Environmental Protection Agency (USEPA). This manuscript has been subjected to review by the National Health and Environmental Effects Research Laboratory's Western Ecology Division and approved for publication. Approval does not signify that the contents reflect the views of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.