Cumulative Distribution Function (CDF) Analysis

Thomas Kincaid

May 23, 2012

Contents

1 Introduction

This document presents cumulative distribution function (CDF) analysis of a GRTS survey design. The resource used in the analysis is small lakes in Florida, which is a finite resource. Lake basins are used to delineate basins among the lakes in the resource. The analysis will include calculation of CDF estimates and testing for difference between CDFs from subpopulations of the resource.

2 Preliminaries

The initial step is to use the library function to load the spsurvey package. After the package is loaded, a message is printed to the R console indicating that the spsurvey package was loaded successfully.

Load the spsurvey package

```
> # Load the spsurvey package
> library(spsurvey)
>
```

```
Version 2.2 of the spsurvey package was loaded successfully.
```

3 Read the survey design and analytical variables data file

The original Florida small lakes data file contains more than 3,800 records and 29 basins. To produce a more manageable number of records, only six basins were retained in the data that will be analyzed, which produced a file containing 930 records.

The next step is to read the data file, which includes both survey design variables and analytical variables. The read.delim function is used to read the tab-delimited file and assign it to a data frame named FL_lakes. The nrow function is used to determine the number of rows in the FL_lakes data frame, and the resulting value is assigned to an object named nr. Finally, the initial six lines and the final six lines in the FL_lakes data frame are printed using the head and tail functions, respectively.

Read the survey design and analytical variables data file

```
> # Read the data file and determine the number of rows in the file
> FL_lakes <- read.delim("FL_lakes.tab")
> nr <- nrow(FL_lakes)
>
```

Display the initial six lines in the data file.

```
> # Display the initial six lines in the data file
> head(FL_lakes)
```

```
         siteID  xcoord   ycoord      wgt    basin           status       TNT
1 FLW03414-0014 8635535 12860896 5.369048 NWFWMD-1          Sampled    Target
2 FLW03414-0046 8636136 12886783 5.369048 NWFWMD-1 Physical Barrier    Target
3 FLW03414-0062 8617834 12869126 5.369048 NWFWMD-1        NonTarget NonTarget
4 FLW03414-0078 8673500 12883071 5.369048 NWFWMD-1 Physical Barrier    Target
5 FLW03414-0086 8631884 12816428 5.369048 NWFWMD-1        NonTarget NonTarget
6 FLW03414-0118 8607699 12856644 5.369048 NWFWMD-1        NonTarget NonTarget
  pH_cat coliform_cat oxygen turbidity
1  (0,6]        (0,5]    9.9       0.4
2   <NA>         <NA>     NA        NA
```

```
3   <NA>          <NA>     NA        NA
4   <NA>          <NA>     NA        NA
5   <NA>          <NA>     NA        NA
6   <NA>          <NA>     NA        NA

>
```

Display the final six lines in the data file.

```
> # Display the final six lines in the data file
> tail(FL_lakes)

           siteID  xcoord   ycoord     wgt    basin            status        TNT
925 FLW03414-3878 8880656 12694963 4.80791 SWFWMD-4               Dry     Target
926 FLW03414-3886 8892406 12732977 4.80791 SWFWMD-4           Sampled     Target
927 FLW03414-3894 8836528 12723056 4.80791 SWFWMD-4               Dry     Target
928 FLW03414-3918 8923107 12725502 4.80791 SWFWMD-4 Landowner Denial     Target
929 FLW03414-3926 8861298 12715824 4.80791 SWFWMD-4               Dry     Target
930 FLW03414-3950 8888601 12715641 4.80791 SWFWMD-4         NonTarget NonTarget
    pH_cat coliform_cat oxygen turbidity
925   <NA>         <NA>     NA        NA
926  (6,8]       (5,50]   1.98       8.2
927   <NA>         <NA>     NA        NA
928   <NA>         <NA>     NA        NA
929   <NA>         <NA>     NA        NA
930   <NA>         <NA>     NA        NA

>
```

The sample of small lakes in Florida is displayed in Figure 1. The sample sites for each basin are displayed using a unique color. First, the levels function is used to extract the set of basin names, and the result is assigned to object basin. Next, the rainbow function is called to select a set of six colors, and the result is assigned to object cols. The plot function is then used to produce the basic figure, but plotting of sample points is suppressed. The for function is used to loop through the set of six basins and plot color-coded points for each basin using the points function. Finally, the legend function is used to add a legend to the figure, and the title function is used to create a figure title.

```
> basins <- levels(FL_lakes$basin)
> cols <- rainbow(6)
> plot(FL_lakes$xcoord, FL_lakes$ycoord, type="n", xlab="x-coordinate",
+      ylab="y-coordinate")
> for(i in 1:6) {
```

```
+    ind <- FL_lakes$basin == basins[i]
+    points(FL_lakes$xcoord[ind], FL_lakes$ycoord[ind], pch=20, cex=0.4,
+           col=cols[i])
+    }
> legend(x="topright", inset=0.05, legend=basins, pch=20, cex=1, col=cols)
> title("Plot of Florida Small Lake Sites Color-Coded by Basin")
```

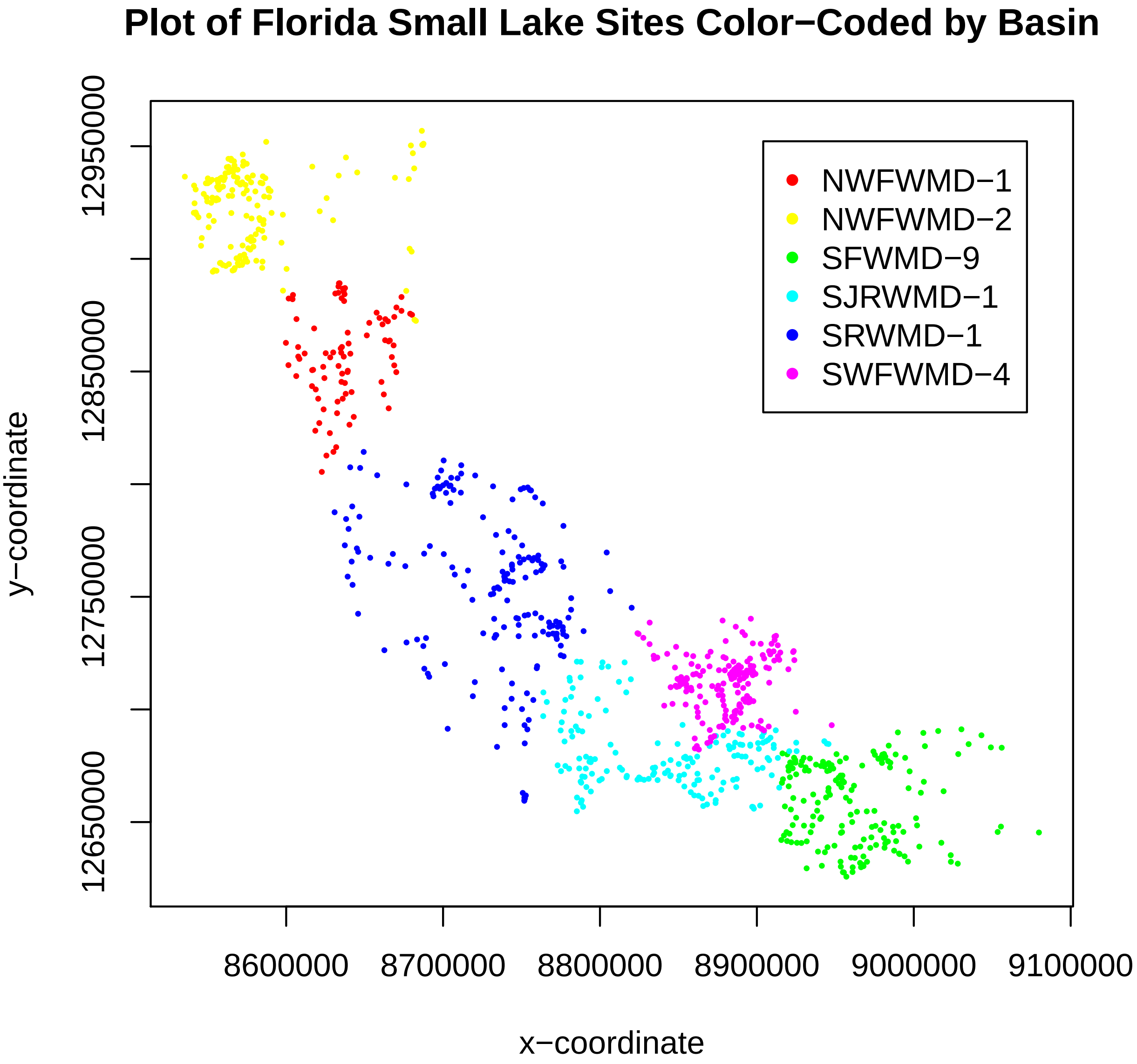

Figure 1: Florida Small Lake Sample Sites.

4 CDF analysis of the dissolved oxygen variable

CDF analysis will be investigated by examining the dissolved oxygen variable. The summary function is used to summarize the data structure of the dissolved oxygen values.

```
> summary(FL_lakes$oxygen)
```

```
Summarize the data structure of the dissolved oxygen variable:
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.830   4.880   6.870   6.468   8.310  12.480     759 
```

Note that there is an extensive number of missing (NA) values. The cont.analysis function will be used to calculate CDF estimates. Four data frames constitute the primary input to the cont.analysis function. The first column (variable) in the four data frames provides the unique identifier (site ID) for each sample site and is used to connect records among the data frames. The siteID variable in the FL_lakes data frame is assigned to the siteID variable in the data frames. The four data frames that will be created are named as follows: sites, subpop, design, and data.cont. The sites data frame identifies sites to use in the analysis and contains two variables: (1) siteID - site ID values and (2) Use - a logical vector indicating which sites to use in the analysis. The Use logical variables in sites is set equal to the value ”Sampled” of the status variable, so that only sampled sites are used in the analysis. The subpop data frame defines populations and, optionally, subpopulations for which estimates are desired. Unlike the sites and design data frames, the subpop data frame can contain an arbitrary number of columns. The first variable in the subpop data frame identifies site ID values and each subsequent variable identifies a type of population, where the variable name is used to identify type. A type variable identifies each site with a character value. If the number of unique values for a type variable is greater than one, then the set of values represent subpopulations of that type. When a type variable consists of a single unique value, then the type does not contain subpopulations. For this analysis, the subpop data frame contains two variables: (1) siteID - site ID values and (2) Basin - which will be used to calculate estimates for each basin individually. The basin variable in the FL_lakes data frame is assigned to the Basin variable in the subpop data frame. The design data frame consists of survey design variables. For the analysis under consideration, the design data frame contains the following variables: (1) siteID - site ID values; (2) wgt - final, adjusted, survey design weights; (3) xcoord - x-coordinates for location; and (4) ycoord - y-coordinates for location. The wgt, xcoord, and ycoord variables in the design data frame are assigned values using variables with the same names in the FL_lakes data frame. Like the subpop data frame, the data.cont data frame can contain an arbitrary number of columns. The first variable in the data.cont data frame identifies site ID values and each subsequent variable identifies a response variable. For this analysis, the response variable is DissolvedOxygen, which is assigned variable oxygen in the FL_lakes data frame.

Create the sites data frame.

```
> sites <- data.frame(siteID=FL_lakes$siteID,
+                     Use=FL_lakes$status == "Sampled")
```

Create the subpop data frame.

```
> subpop <- data.frame(siteID=FL_lakes$siteID,
+                      Basin=FL_lakes$basin)
```

Create the design data frame.

```
> design <- data.frame(siteID=FL_lakes$siteID,
+                      wgt=FL_lakes$wgt,
+                      xcoord=FL_lakes$xcoord,
+                      ycoord=FL_lakes$ycoord)
```

Create the data.cont data frame.

```
> data.cont <- data.frame(siteID=FL_lakes$siteID,
+                         DissolvedOxygen=FL_lakes$oxygen)
```

The frame is a data structure containing spatial location data in addition to other attributes regarding a resource of interest and is used to create a survey design. A frame often takes the form of a shapefile. The frame can be used to obtain size values (e.g., number of lakes) for the populations and subpopulations examined in an analysis. The popsize (population size) argument to cont.analysis provides a mechanism for forcing the total number of lakes for each basin to equal a desired value, i.e., the known number of lakes in the basin. First, the c (combine) function is used to create a named vector of frame size values for each basin. Output from the c function is assigned to an object named framesize. The popsize argument is a list, which is a particular type of R object. The popsize list must include an entry for each population type included in the subpop data frame, i.e., Basin for this analysis. Recall that the Basin population type contains subpopulations, i.e., basins. When a population type contains subpopulations, the entry in the popsize list also is a list. The as.list function is applied to framesize, and the result is assigned to the Basin entry in the popsize list. The popsize argument is included in the call to cont.analysis.

Assign frame size values.

```
> framesize <- c("NWFWMD-1"=451, "NWFWMD-2"=394, "SFWMD-9"=834, "SJRWMD-1"=1216,
+                "SRWMD-1"=1400, "SWFWMD-4"=851)
```

Use the cont.analysis function to calculate CDF estimates for the quantitative variables.

```
> CDF_Estimates <- cont.analysis(sites, subpop, design, data.cont,
+    popsize=list(Basin=as.list(framesize)))
```

The object produced by cont.analysis is a list containing two objects: (1) CDF, a data frame containing the CDF estimates and (2) Pct, a data frame containing percentile estimates plus estimates of population values for mean, variance, and standard deviation. The CDF data frame contains thirteen columns. The first five columns identify the population (Type), subpopulation (Subpopulation), response variable (Indicator), values at which the CDF was estimated (Value), and number of response variable values for each value at which the CDF was estimated(NResp). The next four columns in the CDF data frame provide results for the proportion estimates, i.e., the CDF expressed as percentage values: the proportion estimate (Estimate.P), standard error of the estimate (StdError.P), lower confidence bound (LCB95Pct.P), and upper confidence bound (UCB95Pct.P). Argument conf for cont.analysis allows control of the confidence bound level. The default value for conf is 95, hence the column names for confidence bounds contain the value 95. Supplying a different value to the conf argument will be reflected in the confidence bound names. Confidence bounds are obtained using the standard error and the Normal distribution multiplier corresponding to the confidence level. The final four columns in the data frame provide results for the size (units) estimates, i.e., the CDF expressed in terms of number of lakes: the units estimate (Estimate.U), standard error of the estimate (StdError.U), lower confidence bound (LCB95Pct.U), and upper confidence bound (UCB95Pct.U).

Unlike the CDF data frame, the Pct data frame contains only nine columns since there is a single set of estimates rather than two sets of estimates. In addition, the fourth column is labeled Statistic and identifies either a percentile or the mean, variance, or standard deviation estimate. Finally, since percentile estimates are obtained by inverting the CDF estimate, the percentile estimates do not have a standard error value associated with them.

Use the write.table function to write the CDF estimates as a csv file.

```
> write.table(CDF_Estimates$CDF, file="CDF_Estimates.csv", sep=",",
+             row.names=FALSE)
```

The cont.cdfplot function in spsurvey can be used to produce a PDF file containing plots of the CDF estimates. The primary arguments to cont.cdfplot are a character string containing a name for the PDF file and the CDF data frame in the CDF_Estimates object.

Produce a PDF file containing plots of the CDF estimates.

```
> cont.cdfplot("CDF_Estimates.pdf", CDF_Estimates$CDF)
>
```

The cont.cdftest function in spsurvey can be used to test for statistical difference between the CDFs from subpopulations. For this analysis we will test for statistical difference between the CDFs from the six basins. The cont.cdftest function will test all possible pairs of basins. Arguments to cont.cdftest are the same as arguments to cont.analysis.

```
> CDF_Tests <- cont.cdftest(sites, subpop, design, data.cont,
+    popsize=list(Basin=as.list(framesize)))
```

The print function is used to display results for dissolved oxygen of the statistical tests for difference between CDFs for basins. The object produced by cont.cdftest is a data frame containing eight columns. The first column (Type) identifies the population. The second and third columns (Subpopulation_1 and Subpopulation_2) identify the subpopulations. The fourth column (Indicator) identifies the response variable. Column five contains values of the test statistic. Six test statistics are available, and the default statistic is an F-distribution version of the Wald statistic, which is identified in the data frame as ”Wald-F”. The default statistic is used in this analysis. For further information about the test statistics see the Appendix. Columns six and seven (Degrees_of_Freedom_1 and Degrees_of_Freedom_2) provide the numerator and denominator degrees of freedom for the Wald test. The final column (p_Value) provides the p-value for the test.

```
> # Print results of the statistical tests for difference between CDFs from
> # basins for dissolved oxygen
> options(digits=3)
> print(CDF_Tests)

    Type Subpopulation_1 Subpopulation_2       Indicator Wald_F
1  Basin        NWFWMD-1        NWFWMD-2 DissolvedOxygen  3.125
2  Basin        NWFWMD-1         SFWMD-9 DissolvedOxygen  4.487
3  Basin        NWFWMD-1        SJRWMD-1 DissolvedOxygen 20.300
4  Basin        NWFWMD-1         SRWMD-1 DissolvedOxygen  0.306
5  Basin        NWFWMD-1        SWFWMD-4 DissolvedOxygen 10.675
6  Basin        NWFWMD-2         SFWMD-9 DissolvedOxygen  2.619
7  Basin        NWFWMD-2        SJRWMD-1 DissolvedOxygen  6.072
8  Basin        NWFWMD-2         SRWMD-1 DissolvedOxygen  2.789
9  Basin        NWFWMD-2        SWFWMD-4 DissolvedOxygen  3.835
10 Basin         SFWMD-9        SJRWMD-1 DissolvedOxygen 12.829
11 Basin         SFWMD-9         SRWMD-1 DissolvedOxygen  6.079
12 Basin         SFWMD-9        SWFWMD-4 DissolvedOxygen 14.091
13 Basin        SJRWMD-1         SRWMD-1 DissolvedOxygen 16.915
14 Basin        SJRWMD-1        SWFWMD-4 DissolvedOxygen  5.246
15 Basin         SRWMD-1        SWFWMD-4 DissolvedOxygen  6.398
   Degrees_of_Freedom_1 Degrees_of_Freedom_2  p_Value
1                     2                   55 5.18e-02
2                     2                   57 1.55e-02
3                     2                   57 2.20e-07
4                     2                   54 7.38e-01
5                     2                   51 1.34e-04
6                     2                   55 8.19e-02
7                     2                   55 4.14e-03
8                     2                   52 7.07e-02
9                     2                   49 2.84e-02
10                    2                   57 2.51e-05
11                    2                   54 4.16e-03
12                    2                   51 1.34e-05
13                    2                   54 1.98e-06
14                    2                   51 8.48e-03
15                    2                   48 3.44e-03

>
```

Since there is a fairly large number of CDFs being compared (15), it is reasonable to use a conservative p-value for assessing whether a pair of CDFs is significantly differenct. We will use a p-value no larger than 0.01 to test for significant difference. Using that value, basins with significanly different CDFs are displayed in Figure 2 using "X" to indicate CDFs that are different. Note that only unique pairs of basins are displayed in the Figure 2, which is reflected in the figure by the red line. Significantly different CDFs for pairs of basins located to the right of the red line are displayed in the figure. Nine pairs of CDFs are different: (1) basin NWFWMD-1 is different from basins SJRWMD-1 and SWFWMD-4; (2) basin NWFWMD-2 is different from basin SJRWMD-1; (3) SFWMD-9 is different from basins SJRWMD-1, SRWMD-1, and SWFWMD-4; (4) SJRWMD-1 is different from basins SRWMD-1 and SWFWMD-4; and (5) SRWMD-1 is different from basin SWFWMD-4. The majority of the significantly different CDFs are attributable to two basins. The CDF for basin SJRWMD-1 is different from the CDFs for all of the other basins. The CDF for basin SWFWMD-4 is different from the CDFs for all of the other basins except basin NWFWMD-2. The only other pair of basins with different CDFs are basins SFWMD-9 and SRWMD-1.

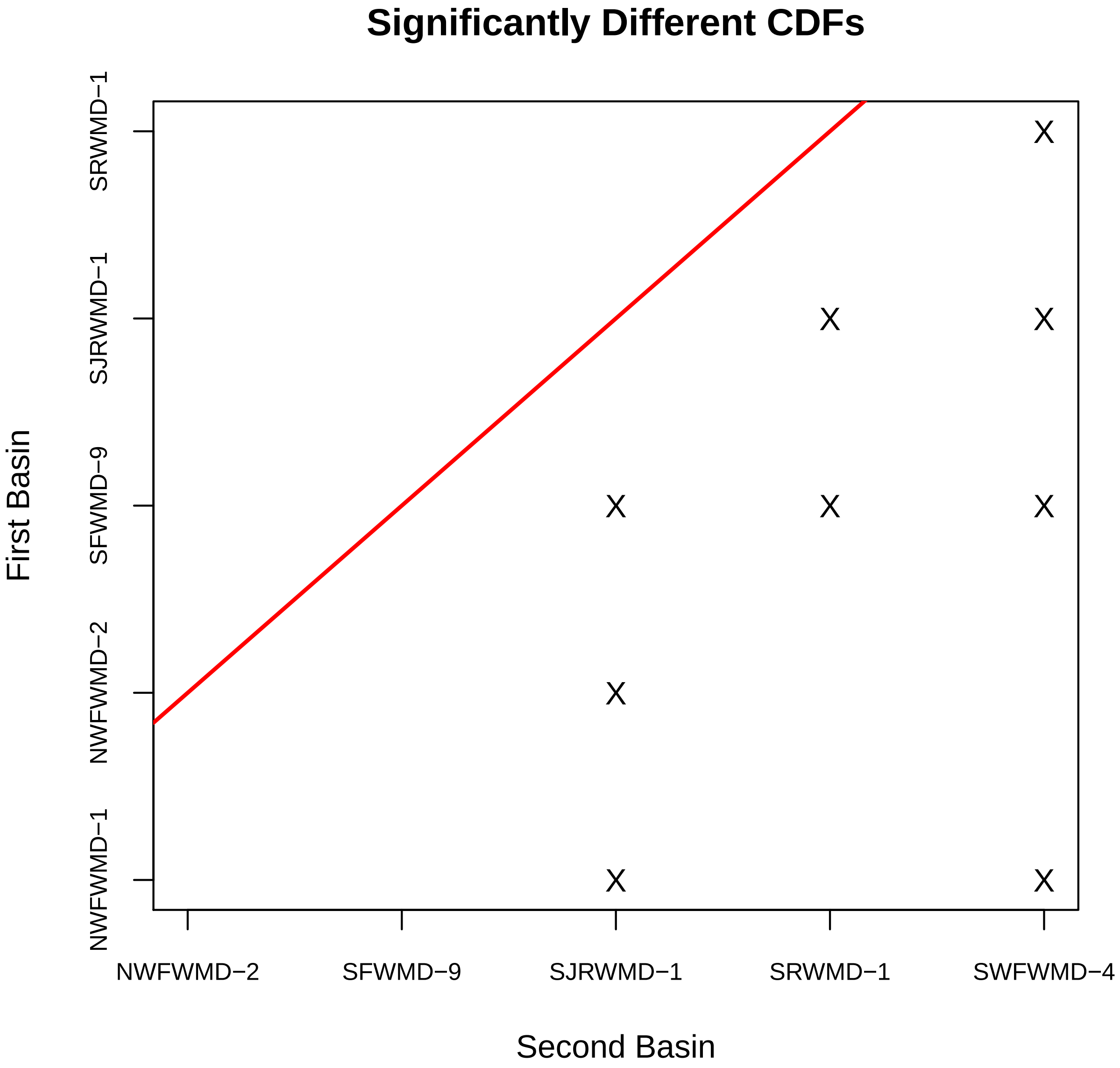

Figure 2: Florida Basins with Significantly Different CDFs.

Use the write.table function to write CDF test results as a csv file.

```
> # Write CDF test results as a csv file
> write.table(CDF_Tests, file="CDF_Tests.csv", sep=",", row.names=FALSE)
>
```

A Appendix – Inferential Procedure

The inferential procedure employed to test for differences between CDFs partitions the two CDFs that are being tested into a discrete set of intervals (classes) and then utilizes procedures that have been developed for analysis of categorical data from probability surveys (Kincaid 2000). The cont.cdftest function calculates the Wald, Rao-Scott first order corrected (mean eigenvalue corrected), and Rao-Scott second order corrected (Satterthwaite corrected) test statistics (Rao and Scott 1981; Wald 1943). Both standard versions of the three statistics, which are distributed as Chi-squared random variables (Rao and Thomas 1988), and alternate version of the statistics, which are distributed as F random variables (Thomas, Roberts, and Singh 1996), are available as options in the cont.cdftest function. The Horvitz-Thompson ratio estimator, i.e., the ratio of two Horvitz-Thompson estimators (Horvitz and Thompson 1952), is used to calculate estimates of the class proportions for the CDFs. Variance estimates for the test statistics are calculated using either the local mean variance estimator or the simple random sampling (SRS) variance estimator. The choice of variance estimator is subject to user control. The SRS variance estimator uses the independent random sample approximation to calculate joint inclusion probabilities.

References

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

Kincaid, T. M. (2000). Testing for differences between cumulative distribution functions from complex environmental sampling surveys. In *Proceeding of the Section on Statistics and the Environment*, Alexandria, VA. American Statistical Association.

Rao, J. N. K. and A. J. Scott (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association 76*, 221–230.

Rao, J. N. K. and D. R. Thomas (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology 18*, 213–269.

Thomas, D. R., G. R. Roberts, and A. C. Singh (1996). Tests of independence on two-way tables under cluster sampling. *International Statistical Review 64*, 295–311.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society 54*, 426–482.