Statistical Consulting Services

3405 Chukar Lane, Clarkston, Washington 99403 (509) 758-8330

Status: Preliminary Report
Date: May 22, 2007
Project Title: Fish Abundance, Biomass, and Condition: 2006 Data, BPA-51

I) Fish Abundance, Biomass, Condition Data

SCS received the initial fish data files from Ryan Hardy, IDFG, on March 21, 2007. The files contained information on site, date, species, sampling effort, fish condition, fish length, and fish weight. A complete listing of these data is provided in Printout #1. A total of 2191 observations were recorded for 6 replications in each of 6 sites. These sites were KR2 (Porthill), KR4 (Shorties Is.), KR6 (Cow Creek), KR9 (Hemlock), KR10 (Yaak Riv.), and KR14 (Wardner, B.C.) From these sites, 15 species were recorded.

Further summary statistics of the data were computed for each site-replication: 1) the number of fish (COUNT), 2) the average fish weight in grams (WT), 3) the average fish length in mm (TL), 4) fish abundance (ABUNDANCE), 5) fish biomass (BIOMASS), and 6) fish condition (K). The values for the last three variables were computed as follows:

ABUNDANCE = COUNT/EFFORT (sec)/3600,
BIOMASS = WT(kg)/EFFORT (sec)/3600, and
$K = (WT/(TL^3)) \times 100{,}000$.

Thus, ABUNDANCE and BIOMASS reflect the number and weight of fish caught in a one hour period, respectively, and K is Fulton's condition factor as outlined in *Blackwell, et al. 2000. Reviews in Fisheries Science, 8(1): 1-44*. These data are given in Printout #2.

As per request of KTOI managers, this report will concentrate on descriptive summarys, trends and sample size determinations for the responses ABUNDANCE, BIOMASS, and fish condition, K. Further analysis of existing or predefined response variables (e.g. growth rate, age, etc.) may be conducted separately and at a future date if deemed appropriate and when the necessary data become available.

II) Summary Statistics

Summary statistics and information for the ABUNDANCE, BIOMASS and K data are presented in Printout #3. Computations were carried out separately for each site over all replications and species.

The first section of Printout #3, for example, gives the ABUNDANCE results for the site KR10. A few of the initial statistics given are the sample size, N = 26, the mean value, 83.45, the sample standard deviation,147.2, and the variance, 21656.3. Other notable values are the skewness, 3.17 and the standard error of the mean, 28.9 Skewness will be discussed in more detail below.

After this initial summary information, measures of location and variability are given. Location measures, such as the mean, relate to where the data is centered. The median, another location measure, is the half-way point in the data when observations are ranked in an ascending or descending order. A third measure, the mode, refers to the most common (frequent) data element. In a data set which was distributed symmetrically about its center, all measures of central tendency would be the same. Variability measures quantify the spread of the data about the center of the data (usually the mean). As seen earlier, the standard deviation and its squared value, the variance, are two such measures. Other variability measures are the difference of the maximum and minimum values or the range and the interquartile range which is the difference between the 75^{th} and 25^{th} percentiles.

The third section titled "Tests for Location" provides tests to assess whether the mean values are equal to zero and do not apply to this data.

Section four, "Tests for Normality" provides a numerical means to assess the closeness of the sample distribution to a theoretical Normal distribution. This consideration becomes important for certain types of analysis. In the case of site KR10, the Shapiro-Wilk test may be used. This statistic should range between zero and one with values closer to one indicating normality. Here, the value of 0.56 is too small relative to 1.0 to indicate Normality. The p-value of 0.0001 confirms this notion, thereby rejecting the hypothesis of Normality for the data.

The fifth and sixth sections provide the percentiles and extreme values for the data. For the site KR10, the data exhibit a large spread of values ranging from a minimum of 5.42 to a maximum of 681.55. This spread is not even, however, as half the data lies below the median value of 24.55. Referring back to the first section of this print out, the skewness statistic of 3.17 also reflects the asymmetry of the data. In symmetric (bell-shaped) distributions the skewness value will be negligible (zero). Here, the positive value of 3.17 indicates that the data has a positive skewness with more small observations than large ones. This is another reason for the aforementioned rejection of theoretical Normal distribution.

The last two parts of each section of the univariate analysis include graphical representations of the data. The first is a stem-and-leaf plot (simple histogram) of the data distribution. The left column of the plot represents the data values while the horizontal bars indicate the frequency of those values. The right hand column gives numeric counts for these frequencies. From this plot, the skewness indicated earlier is obvious with numerous smaller values of abundance and fewer large ones.

The second "Normal Probability Plot" is a graphical assessment of the data distribution (represented by * symbols). The data quantiles are plotted relative to a theoretical normal

distribution (represented by + symbols). If the data followed a normal distribution, the * symbols would lie on top of the + symbols on a 45^{o} line. Here, the data deviate from the line at both the lower and upper ends. This is symptomatic of a skewed data set and reaffirms the results relating to skewness discussed earlier.

Following the univariate analysis of ABUNDANCE, is an analogous univariate summaries for KR10 BIOMASS and K. It should be noted that in this BIOMASS data, the skewness is also large (1.78) and numeric tests and graphs assessing the distribution indicate a skewed distribution for biomass. A similar condition for the distributions of ABUNDANCE and biomass was also found for the other sites, KR2, KR4, KR6, and KR9, as can be seen in the remaining sections of Printout #3. Logarithmic transformation of the data can help mitigate the effects of skewness. Univariate analyses for log(abundance) and log(biomass) are also presented in Printout #3. There it can be seen that this transformation reduces the skewness of abundance and biomass significantly (0.48 and -0.30, respectively). Therefore, this transformed data will be used in the analyses that follow.

III) Trends

In order to maintain consistency with previous fish reports, trend plots, and sample size calculations are based upon the most prevalent (predominant) species. Referring to Printout #1, these species (and the corresponding frequencies) were: LSS (248), MWF (853), NPM (525), PMC (103), RSS (292), and RBT (107). Together, these species accounted for over 97% of the data and the omission of the other species had minimal impact on subsequent computations. Note that similar frequency distributions were observed in analyzing fish data from previous years (see SCS's preliminary fish reports for years 2002 through 2005). Therefore, further analyses of the 2006 fish data will only concentrate on these six species.

Plots of the ABUNDANCE, BIOMASS, and K, trends across sites (recorded as RKM) are given in Printout #4. These plots provide both the mean trend (solid lines) and variability (green boxes with vertical lines) for each site. Each box and vertical line combination represents the mean with its respective minimum and maximum values. As can be seen in the first plot, ABUNDANCE and K show little trend along RKM. The change in mean levels for these responses is small relative to their variability. BIOMASS, however, shows a discernable increase in response levels from RKM 230 to 260 where it then levels off.

Plots of the ABUNDANCE, BIOMASS, and K trends for each species are given in Printout #5. ABUNDANCE, BIOMASS and K show some evidence of trend along RKM with response levels increasing or decreasing, depending on species, across RKM.

Trends for the years 2002 through 2006 are shown overall in Printout #6 and by species in Printout #7. These show that BIOMASS trends were typically similar for these years, however, the trends for K and biomass were variable across years.

IV) Determination of Sample Sizes

The formulation for calculating sample size is:

$$n = (z*s/d)^2$$

where s, d and z are related to the variability, desired precision, and confidence levels, respectively. Due to the skewness of ABUNDANCE and BIOMASS, a logarithmic transformation was carried out on each of these responses prior to estimation in order to meet the necessary distributional requirements of normality.

Estimated sample size requirements for each site are given in Printout #8. Here, z values for the equation above were chosen to provide a range corresponding to 80, 90, 95, and 99% confidence levels. Precision was set at approximately 10% of the overall mean value. The estimated sample sizes for the ABUNDANCE and BIOMASS are small. Note that sample sizes of 1 or 2 are simply indications that the sampling scheme used met or exceeded the specified precision (10% of the mean value, in this case). Thus, the current sampling scheme of 6 replications is sufficient for these responses. Sample size estimates for Condition, K, were more variable. Sample size calculations carried out across species may not be reliable, however, due to species variation in biomass and condition. Including such variability in computations may not accurately reflect appropriate sample size estimates.

In order to counter this effect, sample size estimates were considered separately for each of the 6 species. Using the same precision and confidence levels as above, the sample size estimates were re-computed (see Printout #9). At the 95% level of confidence, the estimated sample sizes for most species were less than 6 with the exception of RBT where higher variability leads to high sample size estimates for ABUNDANCE and RSS shows higher estimated sample sizes for K. Hence, it is important to consider species when estimating sample sizes based on these measurements. Overall, however, the current sampling scheme of 6 replications appears to do well across sites and species.

Note that for all the above calculations, the resulting sample size values are preliminary and also based on limited data. Thus, care should be exercised in applying these results to setting policy regarding future sampling protocols .

V) Additional Remarks

1. The format of the fish data for 2006 was good. Continuation of this trend in data collection should be encouraged as it enhances the quality, reliability, and long term consistency of subsequent analyses.

2. The quality of the 2006 fish data was also excellent. There were no outlying observations noted. Again, this level of quality is to be commended and encouraged for future work.

3. When considered separately by species and across sites, sample size analyses indicate that the current sampling scheme of 6 replications per site is sufficient. This result is consistent with those of previous reports and should be maintained in the future.

5. While some variation in the responses does occur over the years 2002 - 2006, the trend patterns are generally similar with the exception of condition, K, which shows large variation in 2003.

6. As indicated on previous reports, any auxiliary information regarding biological, ecological, environmental, or physical variables could potentially enhance the estimation processes. To be of maximum utility, these variables should be available for each site during all sampling periods. Examples of potentially useful variables might be air and water temperature, thermal or degree day measurements, stream velocity and discharge rates, and habitat information.