STREAM LAB DATA ANALYSIS

We have three goals for this exercise:

1. Introduce you to the use of Excel and Systat for data management and analysis.
2. Develop two graphs and three tables for the Results section of your lab report.
   - Figures: a) habitat and biotic index scores for each site (means and standard error bars)
     b) relationship between a chosen abiotic variable and macroinvertebrate index (correlation scatterplot).
   - Tables: a) water chemistry parameters (means, standard errors, and p-values)
     b) habitat index components (means, standard errors, and p-values)
     c) key biotic index components (means, standard errors, and p-values)
3. Determine which of the measured variables are significantly different among the three sites.

Calculating Means & Standard Errors

You have been given a data file that contains data from all of the groups. You should save this file to your U: drive or desktop before you begin. IF YOU SAVE TO THE DESKTOP, MAKE SURE YOU COPY IT TO YOUR U: DRIVE OR PEN DRIVE WHEN BEFORE YOU LEAVE. There are four worksheets in this file: a) summary sheet, b) water chemistry data, c) habitat index scores, d) habitat index raw values, e) biotic index scores, and e) biotic index raw values. You will do all of your calculations in the data worksheets (b-f), and the information will be transferred to the summary sheet (a) which will be imported into the statistical program Systat for the statistical analysis.

1. To begin, go to the water chemistry data tab. Note that there are rows already labeled mean, std dev and SE (indicating standard error). Standard error is a common way to express the amount of variability around the mean.
2. In the cell where you want the first average to be (C12), type the formula "=AVERAGE(cell range)" where "cell range" encompasses the rows and columns that you want the formula to average.
   - All formulas in Excel start with the "=" sign so if you want to average pH values measured at the Fairhaven Park site and those values are found in column C, rows 2-5, your formula should read "=AVERAGE(C2:C5)".
   - There are several ways to enter the cell range. You can just type it. You can select the range using the mouse (type the formula, open the parentheses and then select the cells you want to include and close parentheses).
3. Your spreadsheet is arranged so that you can copy and paste formulas within a site, rather than having to retype them each time. Highlight the cell where you just entered the formula, select Copy from the Edit menu and then Paste it everywhere you want an average calculated. When you switch sites, you will have to retype formulas to make sure that the cell range is appropriate.
   - This works because Excel uses *relative references* when you copy and paste formulas – if the first formula refers to 8 cells immediately above it, so will any cells you copy the formula into.
   - If you want to copy and paste a formula without changing the cells it references, use an *absolute reference* by adding a "$" in front of the column letter, row number, or both (depending on what you want done) in the original formula before you copy it.
4. To calculate the standard deviation you will use the formula "=STDEV(cell range)". You want to calculate the standard deviation for the same data for which you calculated the average.
5. The standard error is calculated as the standard deviation divided by the square root of the sample size (n, the number of replicate values) (that is, SE = SD/$\sqrt{}$(n-1). Since the formula uses the standard

deviation, it can just reference the cell containing that value. For example if cell C11 contains the standard deviation for the water temperature data, and there are eight values, the formula for standard error would be "=C11/SQRT(8-1)". NOTE: Standard error is the most common measure of variability used in the ecological literature. Because it corrects the standard deviation by sample size, it leads to smaller estimates of variability when we have larger sample sizes. This makes sense, because the more replicates we have, the more certain we can be that our estimate of the mean is close to the true mean for whatever we just measured. You will use standard error when making your graphs and tables.

6. **SAVE your file** (just type CTRL+S). Do this every couple of minutes. It takes less than a second and can save you hours of trying to recover lost work. Make sure that your file is saved to your U: drive or desktop.
7. Repeat these steps until all of the data have been summarized with means, standard deviations, and standard errors.
   - Remember, because of relative references, you don't have to retype the formulas when you move a few rows down to calculate averages and standard deviations for the next site (just copy and paste!).
   - ALWAYS DOUBLE CHECK CELL REFERENCES! When you copy and paste, double click one of the cells to see Excel highlight the cells referenced by that formula. Be careful, if you accidentally click on another cell while you're "editing" a different cell, you'll mess up the formula. ("Undo" was invented for cases like this!) Hit ENTER to get out of the cell edit mode when you've confirmed that the formula is referencing the proper cells.
8. Format the cells with means, standard deviations, and standard errors to have 1 decimal place. Select the range to be formatted, click on the "Home" tab, and select the "Format" menu. Select "Format cells", make sure you're on the "Numbers" tab, then click on "Numbers" in the menu, and select the appropriate number of decimal places. Alternately, right click on highlighted cells and pick "Format cells" from drop-down menu. Go to the "Number" tab, choose the "Number" category and select the appropriate number of decimal places.
9. Repeat the calculations and formatting for the all the variables in the other worksheets.

Hypothesis Testing

In the previous steps, you summarized the data collected by all the groups into two values: the mean and standard error (variability). However, how different do means have to be before you can say that they are fundamentally different? For this we need statistics and hypothesis testing. We will use analysis of variance (ANOVA) and t-tests, both of which statistically evaluate the hypothesis that "the difference in the mean values is simply due to random variation in the data." The value generated is the probability of getting the observed differences in means due to chance alone. Low probabilities (< 0.05) indicate that there's less than a 5% chance that the means are different due to random variation. In other words, the difference between the means is likely to be the result of real differences in the treatments. This probability depends on three things: 1) how different the two means are, 2) how variable the data are, 3) and how extensively you've sampled (i.e., how many replicate measurements you have).

What you must remember is:

IF P≤ 0.05, THE MEANS ARE SIGNIFICANTLY DIFFERENT

IF P> 0.05 THE MEANS ARE NOT SIGNIFICANTLY DIFFERENT

This means that if you find that P=0.06 you must conclude that the means are fundamentally the same, even if one is 9 and the other is 27. If the output indicates that P+0.000, that means that it is a very low

value, which you can indicate in your paper as P<0.001 (that is, less than a one in a thousand chance that differences in means arose solely by chance.

We will be using the statistical software Systat to do our analyses. To do so, we will import our data into Systat to avoid having to retype all the numbers.

1. Make sure your summary data sheet contains all the numbers you need and is formatted with the following restrictions:

a. Only **one** row for the column titles ("headers").

b. Headers should be short but not so cryptic that you can't remember what they refer to. Avoid special characters, as Systat allows only letters and numbers for headers.

c. No blank lines, unless there is a missing value for a particular piece of data. If there is, make sure it isn't the first row.

d. Each row has a site, group, and replicate number.

e. Don't have any numbers in columns with text variables (unless that number is part of a string) and don't have text in columns with number variables.

f. When you're done formatting, SAVE YOUR FILE.

2. **Close your Excel data file.** If you don't, Systat won't be able to open it.

3. Open Systat, then open your data file.

a. Select File…Open…Data, then browse to the location where your Excel data file is.

b. Under "Files of type..", select Excel … .xlsx

c. Select the file and hit "Open". In the subsequent box, hit "OK" to import only the first worksheet (your summary data). Your worksheet should open in Systat. If not, did you remember to close it in Excel?

d. If this method doesn't work, you can copy the data in Excel and paste into Systat, but if you do this, be sure to check that the data are in the correct columns in Systat. Also, you'll need to enter the variable names after copying and pasting the data.

4. ANOVA's. Each variable that we are analyzing has two or three treatments, each with 5-8 replicates . We're combining data to aim for better estimates of stream condition. The treatments are the different sites: SC (Squalicum Creek), PF (Padden Creek in Fairhaven Park), and CA (Chuckanut Creek in Arroyo Park). We want to test if these sites are different from one another in terms of the different variables we measured. When we only have two sites to compare, we use a statistical test called a T-test, specifically designed to compare two means. Think of an Analysis of Variance (ANOVA) as a test that's appropriate for comparing more than two means, so that's what we will use when we are comparing all three streams. We could do three different T-tests, but that approach has a couple of problems: 1) it only uses some of our data points at a time (that is, multiple T-tests are less powerful than ANOVA's for detecting real differences in our treatments), and 2) it breaks some rules about how many tests are appropriate before you need to use a p-value lower than 0.05 for your significance cutoff. We won't go into these statistical details here. Still, many of you may end up needing to use ANOVA's for your independent projects, so we want to introduce them in an appropriate context.

5. Systat. Systat has four main windows, only three of which you will use: 1. an Output Window, whose tab has a .syo extension; and, 2. the Data Window, which has a .syz extension on the tab, and 3. the Graph Window, which will open if you make any graphs in Systat. After you opened your data in step 3 above, the data window will be highlighted. You can click on commands from either the data or output window.

6. To run an ANOVA

a. Click Analyze…Analysis of Variance…Estimate model. Click on SITE$, then Add it to the Factors box. Click on Habitat, then Add it to the Dependent(s) box. Click "OK" to run your analysis.

b. Systat will automatically make a graph of the means and standard errors for you, by site, so you'll be able to see pretty quickly if it looks like there are any significant differences among treatments. As a general rule of thumb, if means don't differ by more than about 2.5 x SE, they probably won't be significantly different.

c. We can see for certain if any are by going to the Output Window. Click on that tab. We will focus on the first three tables in the ANOVA output.

i. The first one just reiterates what your treatments are. Always double check this to make sure Systat analyzed what you think it did.

ii. The second one says what variable you're testing, what the total number of replicates is, and what the R^2 of the analysis it is (Squared Multiple R). R^2 ranges from 0 to 1 and is an estimate of the total proportion of the variability in your data that the different treatments account for (higher is better). An R^2 of 0.99 means that differences among sites account for 99% of the variability in your numbers, and random variation is only 1%.

iii. The third table is where you get the bottom line: are any of the sites different from one another? If the p-value for site is <0.05, then your ANOVA is significant: you know that at least one site differs from one other site. BUT, you don't know which ones differ from which. For that, we need to compare the means.

d. To compare means among sites (do this in all cases when the overall ANOVA gives a P-value of 0.05 or less), click Analyze…Analysis of Variance…Pairwise comparisons. Click SITE$ and add it to the Groups box. Click Tukey under the Test…Equal Variances box, then hit OK.

i. Systat will print another table. The left two rows are the sites you're comparing, "Difference" is the difference between the means of those sites, and p-value tells you whether that difference is significant. Again, if $p < 0.05$, then you can be at least 95% certain that the difference in the means is real (that is, there's less than a 5% chance that the difference just happened from random sampling error). LOW P-VALUES INDICATE SIGNIFICANT DIFFERENCES AMONG SITES. The bigger the difference in the means, the lower the p-value. REMEMBER, you CANNOT say that your sites differ from one another unless the p-value is less than 0.05.

ii. Look at all the comparisons among means to figure out which are different from which. You can use letters to indicate this in your figures (your instructor will describe how).

<u>To Run a t-test</u> (do this for all comparisons for which we only have data for two sites):

a. Go to the analyze tab. Go to hypothesis testing. Select the mean tab. Select two-sample t-test. Select the site as the grouping variable. Select the variable or variables (you can enter all comparisons at once) you wish to compare between sites and add it to the selected variables box. Click ok. To see the output, click .syo tab and use the p-value under separate variance (NOT the one under pooled variance).

7. SAVE your Output and Data files.

8. Do ANOVA's for the overall biotic and habitat index scores and all water chemistry values. In addition, do ANOVA's (or t-tests if only two sites have values) on any components of the biotic or habitat index that are

included as variables used in your figures and tables (see list at top of handout). Report the results in the figure or table, or accompanying legend or the results text (but only report them in ONE of these place).

9. Correlations. You will choose one of your habitat or water chemistry variables as a likely candidate that explains some of the differences in your stream invertebrate multimetric index. This is the beginning of trying to understand the mechanisms underlying the patterns. Note that because we didn't explicitly manipulate the environmental variables, any relationship indicates correlation, not necessarily causation. You will make a scatterplot of the relationship you are testing, then test for the significance of the correlation.

a. Scatterplot. In Systat, select "Graph", then "Scatterplot". Click on the macroinvertebrate index and Add that to the Y-variables. Click on your chosen predictor variable and Add that to the X-variables. Then click on the Smoother tab and select linear. Hit OK to make your graph. This gives you a preliminary look at the data to see if your hypothesis was reasonable. You can play with the graph properties in Systat to make it presentation-quality if you like (look through the other tabs). Once you make the graph you can then save it as a .jpg or .wmf file and insert it into your Word document for your lab report. Alternatively, you can make the graph in Excel (see detailed instructions below). Excel might be better in case you need to edit the graph later and don't have access to a computer with Systat on it. Have you saved your output file recently?

b. Statistical significance? Click the tab that takes you back to the output window (*.syo). Select "Analyze…Correlations…Simple". Choose the two variables that you want to test and Add them to the "Selected variable(s)" window. The following buttons/options should be selected: Type = continuous data, Pearson, Deletion = listwise. Then click the Options tab, check the Probabilities box, and select Uncorrected. Click "OK". Systat will display a graph with histograms of your two selected variables and a scatterplot of the relationship between them. Select the output window (*.syo) and look for the correlation coefficient in the table labeled "Pearson correlation matrix" just above the figure. The p-value is displayed in the Matrix of probabilities, below the figure.

10. When you're done, SAVE your output file, then select everything in your output file and paste it into a Word file (not your lab report), so you can look at it later when finishing your graphs and tables and writing your lab report.

Presenting the Results

Graphs are used to present the most important data (that is, the numbers that answer the biggest level questions that you posed). In this case, we'll graph the mean biotic index scores and the mean habitat index scores. Then we'll make tables for the components of the multimetric indices and for the water chemistry parameters. This is the information that you'll use to try to understand the mechanisms behind any patterns observed for either the habitat or biotic scores. This allows you to "diagnose" what parameters are the major contributors to differences in the indices.

1. Create a new worksheet by right-clicking on one of the tabs at the bottom of the Excel window. Select Insert from the menu and choose worksheet by clicking on the worksheet icon.
2. To make a graph:
    i. Click Insert tab at the top of the spreadsheet.
    - Select "Column" under chart type, and click on the top left box for chart sub-type. Your graph should appear in the next window.

- Right-click on the blank chart and select “Select data” from the drop-down menu. Select the “Add” tab, then place your curser in ”series values” tab. Make sure the box only has the “=” sign then click on the Habitat Index Scores worksheet tab to go to that page, click the cell that has the mean habitat score for the Squalicum Creek site. If you hold in the ctrl button, you can click multiple cells to have the means from all three sites plotted. Click ok.
- To add the mean Biotic Index Scores to the plot, click “Add” to add a new data series, this time selecting the means from the Biotic Index Scores worksheet.
- At this point you can name your Series. Click on the series, and select “Edit.”Enter names into the the “Series name” box. Series 1 is the Habitat Index Scores data and series 2 is the Biotic Index Scores data. Type the labels you wish to have for your x-axis into cells A1:A3. Right-click on your graph to get back to the “Select data” menu. Then you can add them to the graph by clicking the “Horizontal (category) axis labels” box and highlighting cells A1:A3 for the range.
- Add X- and Y- axis labels. Click on your graph, then the “Layout” tab. Choose then “Axis titles” icon. Choose “rotated title” for Y-Axis. Don’t add a graph title! You will be writing our own figure caption in Word. You should also remove the gridlines by clicking on them and deleting. Add the chart as an object in your summary page. Voilá! You’ve done it. Make sure to save your work.

3. To add error bars, go to the “Layout” tab, click “Error bars” menu, then the “error bars” tab. Click on “More error bar options” and go to the “custom positive ” box. Go to the worksheet and select the range of cells for the PC habitat and biotic SE’s. Go to the “custom negative“ box and select the same cells. Then repeat the whole procedure for the other metric, making sure to select the appropriate cells.
4. Remove the gray background to the graph and modify the colors of your bars so they are easily legible in black and white (still the format of most scientific papers). Light gray and dark gray (or black) bars work pretty well. Just double click on the appropriate area and a “properties” menu will appear, from which you can modify many of aspects of your graph.
5. To remove the extraneous decimal places on the Y-axis labels, right-click the axis label, choose “Format axis” from the drop-down menu, and select “Number,” and type “0” into the decimal place box.
6. Make a graph of your correlation data. Go to the graph page in your Excel spreadsheet and select the “Insert” tab.
    - Choose Scatter from the Chart type menu, using the style in the top left window (no added lines). Hit “Next”.
    - Right-click on the blank chart and select “Select data” from the drop-down menu. Select the “Add” tab, Click on the Summary data sheet and select the range of cells for your X (predictor) variable. This will be the habitat index or one of the stream variables. Then click in the Y Values box, select the Summary data sheet and select the range of cells for your Y (response) variable. This will be the biotic index scores for each group at each site.
    - Don’t add a title, but do add X- and Y-axis labels (refer to above directions for adding X- and Y- axis labels) and get rid of the gridlines and the legend. Hit “Next”, then add the graph as an object in your worksheet. Hit finish.
    - Right click on the data points, select “Add trendline”, and select the “linear” box. Choose the options to display equation and R-squared. Don’t include equation and R-squared value on your final figure, but write down the information for your caption. Hit “Close”.
    - Get rid of the grey background in your chart. Have you saved your worksheet recently?
7. You can copy and paste these graphs into a Word file for your report. There you can add figure captions BELOW the graphs. These legends should have a title for the graph and any necessary

explanatory notes so the reader can interpret the graph without having to read the text of your results section. You can also add letters (a, b, c) to indicate significant differences in the bar graph: any bars that have the same letter aren't significantly different from one another.

8. Make the tables. You will be making three tables: one with the components from the habitat index (but not the final multimetric values), a second with the components of the biotic index (but not the final multimetric values), and a third with the water chemistry data.
   - Return to the Water Chemistry worksheet, and save a copy of the worksheet. Open the worksheet containing the figure you just made, and type in the site descriptions and column labels from the Water Chemistry worksheet (with units). Type in the mean ± standard error (you can find the "±" symbol in the "Insert…Symbol" menu.) into each cell of the table.
   - Make sure that you accurately transfer values into your table, and think about whether the numbers make sense. For example, if your worksheet indicates a mean pH value of 17.6 there is something wrong with the calculations!
   - You should present two decimal places for each variable and three for p-values. Make sure to include leading zeros with all numbers less than one.
   - Repeat these steps to create a table of the scores for the habitat index components (also including stream width and stream depth) and another with mean and SE of the actual values (NOT the metric scores) for %Chironomids, %EPT, total taxa, and total individuals.
   - When your tables are constructed, you can copy them into a blank Word document – be sure to leave a few empty rows in the document and between each table before you paste.
   - In Word, add a table legend ABOVE the table. A legend should have a title explaining what data are shown in the table, descriptions of any abbreviations, and any other notes (e.g., for significant differences among means) necessary for the reader to interpret your table without having to refer to the text of your results section. You can use superscripted letters to indicate significantly different means, as with a figure.
   - You will need to format the tables to make them look good in Word – SAVE OFTEN.
     - If necessary, you can adjust column widths by clicking on the lines and dragging them.
     - You can also add line separators to set off the headings and site labels from the numbers by using the "Format…Borders and Shading" menu. Typically, tables have a double line separating heading from the rest of the table and no lines separating the individual numbers.

     - AT THE END OF THE DAY YOU SHOULD HAVE::
     - 2 Figures

       - Habitat and Biotic indices

       - Correlation between Biotic index scores and a potential contributor
     - 3 Tables
       1. Water Chemistry
       2. Habitat index components (all), plus stream width and stream depth.
       3. Biotic index components: % EPT, % chironomids, total taxa, total individuals
     - A complete set of statistical values (p-values, R^2-values, etc.) in Word form so you can refer to it later.

What does it all mean?

- How do the indices (biotic and habitat) at the two sites compare? Are they significantly different?
- If there are differences in biotic indices, what factors appear to drive those differences? How about for the habitat indices?
- Was the correlation between your predictor variable and invertebrate index significant? What does this tell you about potential mechanisms influencing invertebrate community composition?
- Similarly, if there are no differences, why might that be? Remember that insects integrate a lot of environmental variation over a long time, whereas the Eureka measurements were only a single snapshot in time.
- How might the water chemistry, or specific habitat variable, results be affecting the habitat and/or biotic indices?
- Were the streams similar to each other in terms of the control variable (width, depth, and velocity)? If not, how might this influence your results? Are there other important variables that we did not measure, but that could influence the insect community?
- Are there any other interesting ways you might look at the data?