**Environmental Data Laboratory**
**Professor Kirchner**

# Laboratory 8: t-tests and ANOVA in JMP

## Part A. Laboratory intercomparison study--dissolved oxygen measurements

Environmental monitoring data often come from laboratory measurements of chemical concentrations in air, soil, or water. How can one verify the accuracy of such measurements? One way is through intercomparison studies, in which identical samples--of known composition--are analyzed a number of labs. The analyses are performed "blind"; that is, the known concentrations are kept secret from the labs analyzing them.

The file "dissolved oxygen.jmp" contains the results from one such intercomparison study (Wilcock, Stevenson, and Roberts, *Water Research* 15, 321-325). Open that file now. Dissolved oxygen is an important water quality variable, one with direct significance to aquatic organisms; concentrations of at least 5 mg/l are usually needed to maintain healthy fish populations in natural waters. In this study, two samples with known dissolved oxygen (or DO for short) concentrations--1.20 mg/l and 5.86 mg/l--were sent to 39 participating labs. Many of these labs used the Winkler (titration) method, but several also used membrane electrode methods. Results from these analyses are given in four columns, labeled with the known concentration of the standard ("std 1.2" or "std 5.86") and the method used ("Winkler" or "electrode").

So how well did the labs do? Look first at the measurements made by the Winkler method on the 1.2 mg/l standard. Using the "Distribution" platform, display the distribution of the values reported by the labs. •1: Does the distribution look like it's centered around the true value of 1.2 mg/l? What's the mean of the measurements? What is the uncertainty in the mean, as expressed by its standard error?

One way to more precisely test for bias in the measurements is by asking, "if the labs were unbiased--and thus the long-term average of their readings were 1.2 mg/l--what are the chances of getting a set of measurements that deviates from 1.2 mg/l by this much?" This can be tested statistically using the one-sample t-test or a comparable nonparametric test. Here's how it's done in JMP. Click on the small triangle at the top of the histogram, and select "test mean...". Then enter the value you would like to compare with (in this case, 1.2). There's also an option to also run a nonparametric test (Wilcoxon signed-rank test); this is always a good idea, since the results from such a test will generally be much less sensitive to the shape of the distribution of the data. JMP shows the results of such tests as a set of three p-values: p>|t| is the significance level (alpha) associated with a two-tailed test, and p>t and p<t are the significance levels associated with each of the possible one-tailed tests (greater than the value you typed in, and less than that value, respectively). •2: Assuming a two-tailed test (since we're interested in deviations from the true value in either direction), what is the significance level (p-value) of the t-test? The Wilcoxon test? What do these results imply about the possibility of persistent bias in the measurements? •3: The outcomes of the two tests don't differ by much in this example, but if they did, can you think of reasons to prefer one or the other?

•4: Now, repeat steps •1 and •2 for the other three data sets, recording the means and their uncertainties, and the significance levels of their deviations from the known standard concentrations. Note that the concentrations in the samples are below the equilibrium concentration in contact with normal air; thus, samples exposed to air will become contaminated with oxygen over time; this might explain some of the deviation from the standards.

Since both measurement methods (Winkler and electrode) are commonly used, and since measurements made with either method are often compared (e.g. to infer whether DO levels have changed over time) it is critical to know whether the two methods give comparable results. That is, when measured on the same samples, do the two methods yield the same results? This can be tested here, because some labs used both measurements. Thus you can do *paired comparisons* to test for bias between the two methods. Construct a column that calculates the difference between Winkler measurements and electrode measurements on the 1.2 mg/l standard, and another column that does the same thing for measurements of the 5.86 mg/l standard. If the two methods give comparable results, then these differences should not be distinguishable from zero. •5: Test that hypothesis, and answer the following for both data sets: a) How big is the deviation from perfect agreement between the two methods (what's the mean of your column of differences)? b) How uncertain is it (what's its standard error)? And c) How unlikely is that deviation to arise by chance? Does it make a big difference whether you evaluate that likelihood parametrically or nonparametrically?

## Part B. Analysis of Variance--Mercury concentrations in periphyton, South River, Virginia

(This section assumes that you've read chapter 7 in the JMP manual. If you haven't, you should do so now.)

Some chemical contaminants bioaccumulate so efficiently that the best way to look for them is not to sample the water itself, but to sample aquatic organisms instead. The file "periphyton Hg data.jmp" contains a small set of mercury concentrations, measured in periphyton at six locations along the South River, Virginia, above and below a mercury contamination site (Walpole and Myers, 1985). Periphyton live on rooted aquatic plants; this makes them ideal for this kind of study since they are unlikely to travel from one site to another.

The question at hand is whether mercury concentrations increase downstream, reflecting contamination of the river. The mercury concentrations were measured at six fixed stations, numbered 1 through 6 in the downstream direction. Concentrations at all six stations were measured on the same six days, also numbered 1 through 6. The day-to-day differences are trivial compared to the station-to-station differences, so they won't be considered further here. Using the "fit Y by X" platform, plot the Hg concentration (Y) as a function of location (X). Note that since the station number is a nominal scale variable, it does not appear on a continuous axis.

One way to check for changes in mercury concentrations downstream would be to compare the concentrations of each station to each other station, using a t-test. This kind of unplanned multiple comparison is generally a *bad idea*, since it can significantly inflate the overall risk of "false positives". For example, among these 6 sites there are 15 unique site-to-site pairings; if the risk of a "false positive" in any one comparison is $\alpha_c$=0.05, then the risk of *one or more* such false positives in 15 pairings is $\alpha_e=1-(1-\alpha_c)^{15}\approx 0.54$, *not* 0.05.

The right way to compare several sites simultaneously is either through Analysis of Variance (ANOVA for short), or a non-parametric counterpart such as the Kruskal-Wallis test. If ANOVA is performed, it can be followed by the Tukey HSD ("Honestly Significant Difference") test, which can determine which individual sites differ from each other, while controlling the overall "false positive" risk to the desired level.

However, ANOVA assumes that measurements at each site are normally distributed, and that the variance is homogeneous (that is, the amount of scatter at each site is approximately the same). You can check for this by using the "Quantiles" option and the "display" >> "std dev lines" option, both under the "oneway analysis..." triangle (you should <u>not</u> use the "means diamonds" for this purpose, since they are calculated from pooled variance and therefore assume that the scatter is the same at each site). •6: Looking at the mercury concentrations from station to station, does it appear that the conditions for ANOVA are even approximately met?

If the conditions for ANOVA are not met, there are two alternatives. One is to <u>transform</u> the data until they conform to the conditions, and the other is to use nonparametric methods with less restrictive conditions. Let's try the first approach here. Recall the "ladder of powers" from the data transformations toolkit:

| | power | | transformation |
|---|---|---|---|
| | 3 | $x^3$ | cube |
| | 2 | $x^2$ | square |
| start here-------> | 1 | $x^1$ | identity (no transformation) |
| | 1/2 | $x^{0.5}$ | square root |
| | 1/3 | $x^{1/3}$ | cube root |
| | 0 | log(x) | logarithmic (holds the place of zero) |
| | -1/2 | $-1/x^{0.5}$ | reciprocal root |
| | -1 | $-1/x$ | reciprocal |
| | -2 | $-1/x^2$ | reciprocal square |

The ladder of powers says that if your data are positively skewed (long upper tails) and/or the variance increases as the mean increases, transform <u>down</u> the ladder of powers (that is, to square roots, cube roots, logs, etc.). If your data have negative skew, and/or the variance decreases as the mean increases, transform <u>up</u> the ladder of powers.

Your data clearly are begging to be transformed down the ladder of powers. The log transform is convenient because it's easy to interpret: a difference of *q* log units between two groups implies that the un-transformed values in the two groups differ by a multiplicative factor of $10^q$. Since the log transform has worked in the past, try it here. Create a column that calculates the logs of the concentrations, and plot those as a function of site. Check the scatter at each site using the "quantile boxes" option under the "display" triangle. Now which way do you need to move on the ladder of powers? Up, right? So try a square root transform, and plot that next.

Now does the variance at each site look approximately equal (at least so there isn't any clear pattern of increase or decrease with an increasing or decreasing mean)? You can (and should) evaluate this formally, using the "unequal variances" option in the "oneway analysis" pull-down triangle. This shows the results from four different ways of testing whether the scatter in each group is the same. The basis for each of these is explained briefly on p. 147 of the JMP manual. Only the Bartlett test shows a statistically significant deviation from homogeneous variance; although this test is the most powerful of the four, it is also very sensitive to departures from normality, so we might be justified in deciding to ignore it here. At the bottom of the window is a box that shows the Welch ANOVA, which, like Welch's approximate t-test, is designed to perform well even if the variances at each site are unequal. If results from this test disagreed with a "normal" ANOVA, then it would become critically important to figure out which one was right (and thus whether the variances were in fact equal).

Next, using the "means/ANOVA/t-test" option from the "Analysis" triangle, test to see whether there are significant differences among the different means. For help in interpreting the output from this test, consult Chapter 7 of the JMP manual and Chapter 5 of *The Statistical Sleuth*. •7: do the ANOVA results suggest that there are significant differences among the means at the various sites? Why or why not?

Of course, it would be helpful to be able to localize these differences among the sites. That is, it would be useful to decide which sites are statistically distinct from each other, and which are statistically indistinguishable. The proper way to do this, as explained above, is *not* with repeated application of the t-test, but instead is Tukey's HSD test. JMP does this automatically for you when you select "compare means" >> "all pairs, Tukey HSD" from the "oneway analysis" triangle, so do that now.

The results from Tukey's HSD test appear in two forms. One is a table that shows the differences between the means of each pair of sites, minus the least significant difference (which as been adjusted for the number of different comparisons being made, to keep the overall "false positive" rate at alpha). Thus positive values in this table indicate sites that are "significantly" different from one another, and negative values indicate sites that are indistinguishable.

The other useful output is the set of comparison circles next to the x-y plot. These are centered on each of the group means, and their radii are proportional to the uncertainties in the means. For more information about how to interpret the comparison circles, consult the JMP manual or use the help feature. The comparison circles are easy to use: click on any one, and all the others that are not statistically different will highlight as well.

•8: Which sites are distinct from site 1 (which is upstream of the contamination site), and which are indistinguishable from it? •9: Which sites are distinct from site 3, and which are statistically indistinguishable? •10: The mean concentration decreases between sites 5 and 6, breaking the pattern formed by the rest of the data. Is this something that demands an explanation, or could it just be statistical noise?

## Part C: Analysis of Variance--Acid mine drainage

Now try one with a little bit less hand-holding. Acid mine drainage is formed when water percolates into mines and oxidizes sulfide ores, generating sulfuric acid, which in turn mobilizes many different metals, including iron. In coal-mining districts of the northeast, iron oxide from acid mine drainage stains many streams a brilliant color of day-glow orange (including several streams Prof. Kirchner used to swim in as a kid--which might explain a few things, but that's another story). Reclamation projects (including sealing mine shafts and re-contouring and re-vegetating open-pit mines) have been used to reduce the severity of acid mine drainage.

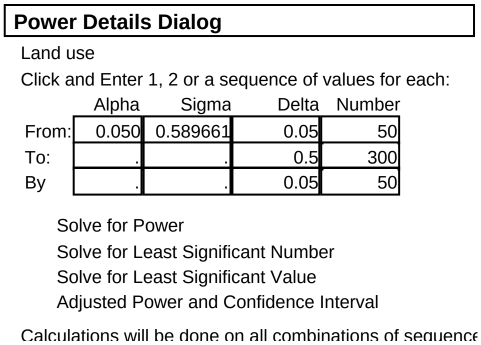
Open the data file, "acid mine drainage.jmp". This file contains measurements of iron concentrations in 120 streams in coal-mining districts of Ohio, grouped according to land use in their basins: unmined, reclaimed (mined and later reclaimed), and abandoned (mined but never reclaimed). For comparison, federal drinking water standards for color and taste permit iron concentrations up to 0.3 mg/l in drinking water. Your task is to evaluate the effects of land use (mining and site reclamation) on iron concentrations in streamwater. One might expect that mining hurts water quality (that is, concentrations will be higher at abandoned mines than in unmined basins), that reclamation helps water quality (that is, concentrations in reclaimed basins will be lower than at abandoned mine sites), and that reclamation is not a complete cure (that is, concentrations in reclaimed basins will be higher than in unmined basins).

•11: Decide whether the data can be used in ANOVA "as is", or whether they will need to be transformed. If so, find an appropriate transformation. Explain the rationale for your choice.

•12: If you can validly do so, perform an analysis of variance, and use Tukey's Honestly Significant Difference to determine which of the categories of streams are statistically distinguishable from each other. Briefly explain what you did and why, and what results you got.

Note that you could also approach this analysis with the three specific pairwise hypotheses outlined above. Testing each of these hypotheses would not violate the prohibition on multiple comparisons, as long as the hypotheses were developed before one had seen the data. You can test any two land use types directly against one another by excluding the third and using the "means Anova/t-test" option under the "analysis" triangle. A two-sample (unpaired) t-test is mathematically equivalent to a one-way analysis of variance with two groups, so many statistical packages calculate them that way. Note that you *should not* calculate the differences between individual streams (in a paired comparison) because there is no intrisic pairing among them; that is, there's no relationship between the first reclaimed stream and the first mined stream and so forth. •13: try a t-test for each of the three specific hypotheses outlined above, and for each test write down the effect size (the estimate of the difference between the two groups), with its 95% confidence interval. If you transformed the concentrations, also show these estimates transformed back into the original units. Also write down the statistical significance level of the difference, but note that since the hypotheses above are *one-tailed*, the actual alpha is half of the two-tailed alpha (the "prob>|t|") reported here.

Finally, note that you can use JMP to do quick, convenient power analysis. For example, in the ANOVA/t-test you just did, in which you compared the unmined and reclaimed areas, click on the triangle next to the "one way ANOVA" button and select "power". The power details dialog allows you to specify any combination or range of alpha (significance level), sigma (pooled standard deviation), delta (difference between groups) and number (number of measurements in all groups, here assumed to be split evenly among them). To check the power of several possible tests, you would pull down "power" and fill out the Power Details Dialog. For example, to evaluate the power to detect differences ranging from delta=0.05 to delta=0.5, using sample sizes from 50 to 300, with sigma and alpha held constant, you would fill out the power details dialog as follows, with the "Solve for Power" box checked (obviously, the value for sigma will depend on the particular transformations you have used):



**Power Details Dialog**

Land use

Click and Enter 1, 2 or a sequence of values for each:

| | Alpha | Sigma | Delta | Number |
|---|---|---|---|---|
| From: | 0.050 | 0.589661 | 0.05 | 50 |
| To: | . | . | 0.5 | 300 |
| By | . | . | 0.05 | 50 |

Solve for Power
Solve for Least Significant Number
Solve for Least Significant Value
Adjusted Power and Confidence Interval

Calculations will be done on all combinations of sequence

•14: Using the power details dialog, test how power changes with effect size (delta) and sample size (Number) across the range specified above. Given a sample size of 50, what is the smallest difference that could be detected with 90% reliability (that is, power of 90%)? With 99% power? Given a sample size of 300, what is the smallest difference that could be detected with 90% power? With 99% power?