EXPERIMENT 3

Density: A Study in Precision and Accuracy (Part B)

Validity

In this part of the statistics lab, we will consider two additional aspects of statistical treatment of data which are very important to scientists. The first is validity. **Validity** is a measure of how well cause and effect are correlated. Testing claims of the validity of a cause and effect relationship between two variables is perhaps the most basic part of what scientists do. If the effect of a new drug in treating a particular disease is being tested, the drug must be tested on a set of patients as well as a control group which receives a "placebo." The obvious question is whether or not there is a significant difference in symptoms of disease between those who took the drug and those who were given a placebo.

Validity is a measure of whether two different results are truly different statistically. For example, a scientist could study the colon cancer rates of those who eat Wheaties and those who do not. Let us imagine that the colon cancer rate of those who eat Wheaties is 24.5 per thousand, while those who do not eat Wheaties have a cancer rate of 24.0 per thousand. Is the scientist justified in reporting that eating Wheaties can increase your likelihood of getting cancer? The answer is almost certainly no!!! The two different results almost certainly do not differ enough to statistically justify concluding there is a relationship between eating Wheaties and getting colon cancer.

The problem of determining validity of a result is especially difficult in the biological sciences, and even more so in the medical sciences. For example, consider the following hypothetical study. A group of subjects was surveyed and it was discovered that people in the army have a 30% higher lung cancer rate than those not in the army. This 30% difference is certainly statistically valid. Conclusion: being in the army causes lung cancer. Wrong!!! What this study fails to do is to adjust the results for smokers. In fact, those in the army have a 40% higher rate of smoking. It was not being in the army which caused cancer, it was smoking.

The conclusion is that anyone doing a scientific study must very carefully consider all the relevant variables which could conceivably effect a given result. Once all the variable have been controlled for, the results must still be checked for statistical validity. In other words, is there a valid correlation between a change in a given variable and the result measured.

The *t* Test

The most common statistical test for whether a scientific measurement of an effect is valid is the *t* test. For a given set of data, one being the test, the other being the control, the question is whether the average value measured is statistically different. Is there a valid effect? To provide an example, consider the following data:

TABLE 3.1

measurement #	height of plant using just water	height of plant using "Mighty Grow"
1	58 cm	64 cm
2	62 cm	55 cm
3	53 cm	58 cm
4	61 cm	66 cm
5	54 cm	56 cm
6	57 cm	62 cm
average	57.5 cm	60.2 cm

Conclusion: "Mighty Grow" makes the plants grow faster. Not so fast! We must apply the *t* test. Look at Equation 3.1

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p}\sqrt{\frac{N_1 N_2}{N_1 + N_2}} \qquad \textbf{(EQ 3.1)}$$

Where $\bar{x}_1$ is the average of the first set of date, $\bar{x}_2$ is the average of the second set of data, N_1 and N_2 are the number of measurements for each set of data, and s_p is the pooled standard deviation of the two sets of data. The pooled standard deviation is given by Equation 3.2.

$$s_p = \sqrt{\frac{\sum(x_{i1} - \bar{x}_1)^2 + \sum(x_{i2} - \bar{x}_2)^2}{N_1 + N_2 - 2}} \qquad \textbf{(EQ 3.2)}$$

The value of *t* is calculated and compared to a *t* table. If it is greater than the relevant *t* value in the table, then the difference between the two measurements is valid. A table of *t* values is included.

For example, from the data in Table 3.1 on page 22, one can calculate the *t* value to be:

TABLE 3.2

Set #1 $(x_i - \bar{x}_1)^2$	Set #2 $(x_i - \bar{x}_2)^2$
0.25	14.44
20.25	27.04
20.25	4.84
12.25	33.64
12.25	17.64
0.25	2.24
$\sum(x_i - \bar{x}_1)^2 = 65.50$	$\sum(x_i - \bar{x}_2)^2 = 99.84$

$$s_p = \sqrt{\frac{65.50 + 99.84}{6 + 6 - 2}} = 4.07 \qquad \textbf{(EQ 3.3)}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p}\sqrt{\frac{N_1 N_2}{N_1 + N_2}} = \frac{60.2 - 57.5}{4.07}\sqrt{\frac{6 \cdot 6}{6 + 6}} = 1.14 \qquad \textbf{(EQ 3.4)}$$

Now, checking the *t* table, the number of degrees of freedom is *N*–1. Critical values for t (two-tailed). Use these for the calculations of confidence intervals. For example, use the 0.05 column for the 95% confidence interval.

TABLE 3.3

degrees of freedom	0.50	0.10	0.05	0.02	0.01
1	1.000	6.314	12.706	31.821	63.657
2	0.816	2.920	4.303	6.965	9.925
3	0.765	2.353	3.182	4.541	5.841
4	0.741	2.132	2.776	3.747	4.604
5	0.727	2.015	2.571	3.365	4.032
6	0.718	1.943	2.447	3.143	3.707
7	0.711	1.895	2.365	2.998	3.499
8	0.706	1.860	2.306	2.896	3.355
9	0.703	1.833	2.262	2.821	3.250
10	0.700	1.812	2.228	2.764	3.169

Since we had six measurements, it is five. At the 90% confidence level, with 5 degrees of freedom, *t* = 2.015. Since our value for *t* was 1.14, there is not a valid correlation between use of the fertilizer and plant height. If we had used the 50% confidence level, *t* = 0.727, and the result would be valid. In other words, at a 50% confidence level, there is at least a small statistical effect of using the fertilizer.

The “Experiment”

Perform a *t* test to see if there is a statistically valid relationship between number of bean seeds sprouted per 100 and exposure to UV light.

TABLE 3.4

Experiment #	**# sprouted without UV irradiation**	**# sprouted with UV irradiation**
1	87	71
2	72	64
3	88	80
4	81	69
5	69	70
6	78	70
7	80	72
Average		

1. Calculate the average for each set of data and fill in the blank in the table above. Then do the calculations to find *t* for the two sets of data. Show your calculations in your lab book.
2. What is the number of "degrees of freedom" for the data above?
3. Find the value of t from the attached *t* table using your number of degrees of freedom at the 90% confidence level. Compare to your *t* calculated above. According to your result, is there a significant difference in the seed-sprouting rate for seeds irradiated with UV light? If not, is the difference significant at the 50% confidence level? If yes to the 90% confidence level, what is the highest confidence level at which the result is valid according to the table?

Least Squares Analysis Of Data

Least squares analysis of data is a statistical method for determining the best fit straight line to a set of data. There is hardly any more common thing for a chemist to do than to fit a set of data to a straight line, be it in kinetic studies, absorbance/concentration studies and so forth. Chemists almost invariably use a canned program from excel or other software to determine the slope and intercept of the best straight line fit to a set of data. In this experiment, you will actually do a least squares analysis of a set of data by hand. The theory and equations of least squares analysis is provided in an attachment to this lab write-up. You will be doing a simple experiment to determine the density of a solution using least squares analysis of data.

Experiment

Using a 50.0 ml graduated cylinder, measure the mass of the cylinder empty as well as five sets of volume and mass data for the same cylinder and a solution provided. The volumes should be about 10, 20, 30, 40, and 50 ml. Measure both mass and volume with as much precision as the data allows. Record the data in your lab book. That is it!

Calculations

1. Perform a least squares fit to the five pairs of data, assuming that the volume is the independent (*x*-axis) data and the mass is the dependent (*y*-axis) data. Your analysis of the data should include finding both the slope and the intercept (see Equation 3.5 and Equation 3.6), as well as the uncertainty in both numbers (see Equation 3.7, Equation 3.8, and Equation 3.9). Record the slope as *m* = slope ± error in slope and the intercept as *b* = intercept ± error in intercept.
2. What is the physical interpretation of the slope of your graph? Does it agree with the correct answer (look it up) within the uncertainty? What is the %-error? Is a systematic error required to explain your %-error? (explain)
3. What is the physical interpretation of the intercept of your graph? Does this value agree with the correct value within the uncertainty you determined? Calculate your %-error. Is a systematic error required to explain your %-error?
4. Now, make a graph of your data and do a least squares fit to the same data using a canned program such as Vernier or Excel, available on the computers and compare to the values you got by hand.

$$m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \left[\frac{\sum x_i \sum y_i}{n}\right]}{\sum x_i^2 - \left[\frac{\left(\sum x_i\right)^2}{n}\right]} \qquad \textbf{(EQ 3.5)}$$

$$b = \bar{y} - m\bar{x} \qquad \textbf{(EQ 3.6)}$$

$$s_y = \sqrt{\frac{\left[\sum y_1^2 - \frac{\left(\sum y_i\right)^2}{N}\right] - m^2\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}\right]}{N-2}} \qquad \textbf{(EQ 3.7)}$$

$$s_m = \sqrt{\frac{s_y^2}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}}} \qquad \textbf{(EQ 3.8)}$$

$$s_b = s_y \sqrt{\frac{1}{N - \frac{\left(\sum x_i\right)^2}{\sum x_i^2}}} \qquad \textbf{(EQ 3.9)}$$