Experiment 9: Penny Statistics

CH2250: Techniques in Laboratory Chemistry, Plymouth State University

Adapted from "4. Penny Statistics," *Experiments To Accompany Exploring Chemical Analysis, 4th Edition*, Daniel C. Harris (2008), available at http://www.whfreeman.com/exploringchem4e. Originally from T. H. Richardson, *J. Chem. Ed.,* **68:**310 (1991) and R. J. Stolzberg, *J. Chem. Ed*, **75**:1453 (1998).

Useful websites: http://www.gifted.uconn.edu/siegle/research/t-test/t-test.html,

http://www.stat-help.com/spreadsheets.html, http://www.graphpad.com/articles/outlier.htm

http://zimmer.csufresno.edu/~davidz/Stat/LLSTutorial/Formulas/LLSFormulas.html

http://chemistry1.che.georgiasouthern.edu/chem2031/Data/Data1/data1.html

Introduction: Laboratory experiments frequently involve the collection of numerical data, and often as a way to limit the errors, these data are collected in duplicate. Statistical analysis refers to a class of mathematical analyses that can be performed on data to ensure they are the best they can be—that is, they best reflect the true results and give us the best understanding of what happened. This experiment will introduce you to some statistical analysis techniques, particularly those related to analyzing large, repetitive data sets. One such analysis you are certain familiar with: the average, or arithmetic mean.

The analyses to be run include calculating the mean, the standard deviation, and confidence intervals; running *t* tests and least-squares analyses; and identifying discrepant data. A computer spreadsheet (Excel) will be used for all the mathematical analyses.

One of the most useful aspects of spreadsheets, and thus one of the more important things to learn, is the ability to reference cells in equations. Every cell is identified by its column letter and row number (e.g., A1, B10, C55). When you write an equation in a cell, you may simple enter the letter and number of the cell to be referenced, or you may used your mouse to select the cell. Often, you will need to reference a range of cells, which is done by giving the first and last cell, separated by a colon (e.g., A1:A100 for the range of data in the cells A1 to A100), or you may use your mouse to select the cells. If you cut and paste a formula, the pasted formula will reference the cells *relative* to the cells referenced in the original formula (e.g., if the formula in B1 is "A1*2" and this is copied to B2, the formula in B2 will be "A2*2"). Often this is exactly what you want, but sometimes you want a certain reference to always refer to the same cell. In order to do this, use the "$" (e.g., if the formula in B1 is "A1*2" and this is copied to B2, the formula in B2 will be "A1*2").

The data set to be collected consists of the masses of a large number of pennies. U.S. pennies minted after 1982 have a Zn core with a Cu over-layer. Prior to 1982, pennies were made of brass, with a uniform composition (95 wt% Cu / 5 wt% Zn). In 1982, both the heavier brass coins and the lighter zinc coins were made. We will, therefore, only consider pennies made after 1982. In this experiment, your class will weigh many coins and pool the data in order to determine whether the averages mass of pennies each year has changed.

Equipment: None required. If you have a USB mouse, you may wish to bring it to class to facilitate use of the computer.

Below is a table of the functions you will use:

	Name	Formula	Excel Function
1	Number of Observations	n	COUNT(data)
2	Mean	$\bar{x}=\frac{1}{n}(x_1+x_2\cdots x_n)$	AVERAGE(data)
3	Standard Deviation	$s=\sqrt{\frac{\sum_i (x_i-\bar{x})^2}{n-1}}$	STDEV(data)
4	Minimum value		MIN(data)
5	Maximum value		MAX(data)
6	Student's *t*	See Table 4-2 (pg 87)	TINV(α, df)
7	Confidence interval	$\mu_{CL}=\bar{x}\pm\frac{ts}{\sqrt{n}}$	AVERAGE(data) + CONFIDENCE(α, *s*, *n*) AVERAGE(data) - CONFIDENCE(α, *s*, *n*)
8	*t* Test for Comparison of Means	see Equation 4-4, 4-5 in textbook, pg 89	see spreadsheet "8-StatisticalFunctions.xls" on course website*
9	Grubb's test	$G=\frac{\lvert x-\bar{x}\rvert}{s}$	calculate using ABS(), AVERAGE(), and STDEV()
10	Linear Regression (*m* and *b*)	see Equations 4-9, 4-10, 4-11 in textbook, pg 94	SLOPE(data_Y, data_X) INTERCEPT(data_Y, data_X)
11	Standard Deviation of Slope (s_m)	see Equations 4-12, 4-13 in textbook, pg 95	see spreadsheet "8-StatisticalFunctions.xls" on course website*
12	Confidence interval for Slope	$\mu_{m,CI}=m\pm ts_m$	calculate using TINV(α, df) and previously calculated *m* and s_m
13	Standard Error of Linear Regression		STEYX(data_Y, data_X)
14	R-squared		RSQ(data_Y, data_X)
'α' is 1-confidence, thus for a confidence level of 95%, α = 0.05 'df' is the degrees of freedom, *n*-1 *these functions are not implemented directly in Excel, but they can be entered as more complicated functions. This has been done for you in the indicated spreadsheet.			

Procedure:

1. Create a table with the headings of the columns being the years 1983 through 2000.
2. Use an analytical balance to find the masses of about 30 pennies. Record these in your notebook according to their year of minting.
3. Boot-up and log-in to a computer. Start Excel and open a new spreadsheet. Starting in B2 and moving right, enter the years 1983 to 2000 (hint: in B3 enter "B2 + 1". Copy this formula to the remaining cells.)

4. Input the masses of your pennies into the document. Save this spreadsheet! Record the name of the spreadsheet in your notebook.
5. Open a web browser and navigate to http://tinyurl.com/7awm9y7
6. Cut and paste the masses from your spreadsheet into the Google Docs spreadsheet.
7. Once the class has completed entering their data, download the spreadsheet to your computer and open it with Excel. Save the class data spreadsheet with a different name (do no overwrite your original data!) Record the name of the file in your notebook.

<u>Analysis</u>

A. Setting up the Spreadsheet

1. Be sure that the first two rows and first column are blank. If you have already written in the first column, click on the heading for Column A to highlight the entire column, right click on any highlighted cell, and hit "Insert Column." Insert rows in a similar manner.
2. Write your name, the date, and a title for the spreadsheet in cells in the first row.
3. Put the following row headings in the first column: Mean, StdDev, Mean+2.5s, Mean-2.5s.
4. Sort each individual column of data from lightest to heaviest: Highlight the data in just one column, go to the DATA menu, select SORT, and follow the directions that come up. Be sure the year is NOT sorted and that Excel does not select and sort the neighboring data.
5. Set the number format for all the cells that will have data and calculations to have 4 decimal places. Highlight all of these cells, right click, select "Format Cells," be sure "number" is selected, and increase the number of decimals to 4.
6. When you are done, your spreadsheet should look something like Figure 1:

	A	B	C	D	E
1	Dr. Duncan	04/14/11		Penny Statistics Spreadshe	
2					
3		1983	1984	1985	1986
4		2.3081	1.9543	2.3480	2.3324
5		2.3229	2.4343	2.3787	2.4042
6		2.3680	2.4477	2.5075	2.4195
7		2.3746	2.4572	2.5087	2.4666
8		2.4401	2.5193	2.5272	2.4857
9		2.4607	2.5467	2.5338	2.5170
10		2.4638	2.6091	2.5806	2.5744
11		2.5499	2.6631	2.6004	
12		2.5922	2.6792	2.6486	
13		2.6642	2.6847	2.6584	
14		2.6738		2.6715	
15		2.6910			
16	Mean				
17	StdDev				
18	Mean+2.5s				
19	Mean-2.5s				

Figure 1

B. Discrepant Data

1. At the bottom of each column in the appropriately labeled rows, compute the mean and standard deviation (Equations 2 and 3)
2. The Grubb's Test for an Outlier is an excellent way to determine whether any of your data are discrepant and should be thrown out. It should be clear that only the maximum and/or the minimum value in a list would possibly be discrepant. In a cell just to the right of the year 2000 data, find the maximum and minimum masses (equations 4 and 5) of the pennies made in 2000 and then calculate *G* (equation 9) for these values (see Figure 2).
3. Although the Grubb's test is a very rigorous test for outliers, it can be time consuming, because the mean and standard deviation must be recalculated after each discrepant datum is thrown out. Notice that the Grubb's test essentially looks at the number of

1999	2000		
2.3551	2.2861	G value (min)	*Calculated*
2.3859	2.3008	G value (max)	*values here*
2.5150	2.3455		
2.5163	2.3521		
2.5348	2.4169		
2.5414	2.4373		
2.5884	2.4404		
2.6082	2.5257		
2.6566	2.5676		
2.6664	2.6389		
	2.6484		

Figure 2

standard deviations a value is from the mean, and then sets a minimum number of standard deviations that defines a discrepant datum for a given number of observations. To quickly analyze all our sets of data for outliers, we will perform a crude Grubb's test, erring far on the side of caution by choosing our Critical Value to be 2.5 standard deviations. Calculate the range of acceptable data (those within 4 standard deviations) under the data for each year using the formulas (cells "Mean + 2.5s" and "Mean - 2.5s" in Figure 1):

- average + (2.5 × standard deviation)
- average - (2.5 × standard deviation)

4. Analyze your data for grossly discrepant masses (those lying ≥ 2.5 standard deviations from the mean) in any one year. (For example, if one column has an average mass of 3.000 g and a standard deviation of 0.030 g, the 2.5-standard-deviation limit is ± (2.5 × 0.030) = ±0.075 g. A mass that is ≤ 2.925 or ≥ 3.0750 g should be discarded.) Copy and paste the data from each column that is within the acceptable limits to a series of cells somewhat below the original data.
5. Calculate the mean and standard deviation of the new set of data just below it. Be sure to create new row labels for these calculations (Figure 3; note lowest number thrown out from 1984 data).

	A	B	C
1	Dr. Duncan	04/14/11	
2			
3		1983	1984
4		2.3081	1.9543
5		2.3229	2.4343
6		2.3680	2.4477
7		2.3746	2.4572
8		2.4401	2.5193
9		2.4607	2.5467
0		2.4638	2.6091
1		2.5499	2.6631
2		2.5922	2.6792
3		2.6642	2.6847
4		2.6738	
5		2.6910	
6	Mean	2.4924	2.4996
7	StdDev	0.1387	0.2145
8	Mean+2.5s	2.8393	3.0357
9	Mean-2.5s	2.1456	1.9634
0			
1		1983	1984
2		2.3081	
3		2.3229	2.4343
4		2.3680	2.4477
5		2.3746	2.4572
6		2.4401	2.5193
7		2.4607	2.5467
8		2.4638	2.6091
9		2.5499	2.6631
0		2.5922	2.6792
1		2.6642	2.6847
2		2.6738	
3		2.6910	
4	Mean	2.4924	2.5601
5	StdDev	0.1387	0.1022

Figure 3

C. Confidence Intervals

1. Find the year with the highest and the year with the lowest average masses.
2. Label the next row "n" (the number of data points). Calculate "n" (Equation 1) for the data resulting in the highest and lowest average mass.
3. Label the next row "t" (Student's *t*, see Table 4-2, pg 87). Use the TINV(0.05, n-1) function to find the value of Student's *t* at the 95% confidence level for the data resulting in the highest and lowest average mass. Remember that the degrees of freedom = n – 1. (Figure 4) Record these data in your notebook.
4. Calculate the 95% confidence ($\mu_{95\%}$) interval for the highest average mass by hand (Equation 5). Show this work in your notebook.
5. Have Excel calculate the 95% confidence intervals ($\mu_{95\%}$) (Equation 7) for the highest and lowest average masses in the cells below the calculation for *t* (Figure 4). Double check that it found the same interval for the highest mass that you did! Record these intervals in your notebook.
6. Run the *t* Test for Comparison of Means (Equation 8) using the functions prepared for you in the spreadsheet 8-StatisticalFunctions.xls Note the values of *t* stat and P in your notebook.

	1983	1984	1985	1986
	2.3081		2.3480	2.3324
	2.3229	2.4343	2.3787	2.4042
	2.3680	2.4477	2.5075	2.4195
	2.3746	2.4572	2.5087	2.4666
	2.4401	2.5193	2.5272	2.4857
	2.4607	2.5467	2.5338	2.5170
	2.4638	2.6091	2.5806	2.5744
	2.5499	2.6631	2.6004	
	2.5922	2.6792	2.6486	
	2.6642	2.6847	2.6584	
	2.6738		2.6715	
	2.6910			
Mean	2.4924	2.5601	2.5421	2.4571
StdDev	0.1387	0.1022	0.1065	0.0796
	Year of highest average		Year of lowest average	
	1984		1998	
n	9		12	
t	2.3060		2.2010	
mu-95 (max)	2.6387		2.4804	
mu-95 (min)	2.4816		2.3596	

Figure 4

D. Least-Squares Analysis: Do Pennies Have the Same Mass Each Year?

1. Copy all of the data into a single column several to the right of the data. As you copy the data, put the year that corresponds to each mass in the column to the left.
2. At the top and to the right of the data, calculate the number of pennies, average and standard deviation for all the penny masses. Calculate the slope, intercept, standard error of linear regression, and R-sqaured for the best-fit line (linear regression, Equations 8, 11, 12), using the years as the x-data and masses as y-data. Record these values in your notebook (Figure 5).
3. Use the worksheet 8-StatisticalFunctions.xls to find the standard deviation for the slope (Equation 9). Record this number in your notebook.
4. Calculate the 95% Confidence interval for the slope of the best-fit line (Equation 10). Record this in your notebook.
5. Insert an X-Y scatter plot. The x-data are years, and the y-data are the masses of the pennies. Give the graph a title, and label the axes. You do not need to show the legend.
6. While still in the mode to edit the graph, go to "Insert -> Trendline" and insert a linear trendline on the graph.

Year	Mass		
1983	2.3081	n	*Calculated*
1983	2.3229	Average:	*values here*
1983	2.3680	Std Dev:	
1983	2.3746	Slope:	
1983	2.4401	Intercept	
1983	2.4607	Steyx:	
1983	2.4638	R2	
1983	2.5499		
1983	2.5922	mu-95 (max)	
1983	2.6642	mu-95 (min)	
1983	2.6738		
1983	2.6910		
1984	2.3288		
1984	2.3504		
1984	2.4477		
1984	2.4572		
1984	2.5193		
1984	2.5467		
1984	2.6091		
1984	2.6631		
1984	2.6792		
1984	2.6847		
1985	2.3480		
1985	2.3787		

Figure 5

<u>Conclusions</u>

1. Compare the calculated values of *G* (step B2) with the Critical Value of *G* for the given number of observations (see Table 4-4, pg 92 and http://www.graphpad.com/articles/outlier.htm). By the Grubb's test, are these values outliers?
2. Compare the result of the *t* Test for Comparison of Means with the value of Student's *t* you calculated for the largest and smallest average masses of pennies (C4 and C6). Is the difference in these two averages statistically significant (i.e. is the result of the *t* test larger than the larger of the two calculated *t*'s)?
3. The P you calculated in C6 is a statistical function that tells us the chance of the differences between the means of two groups being due to random chance. Say you get a P value of 0.10 (or 10%). This means that there is a 10% chance that the differences between your two groups are due to random chance alone. Another way to say this is that there is a 90% chance that the differences between these two groups is significant. Normally will say that a P value of .05 or less is significant. What does the P value you calculated tell you about the significance of the difference between the largest and smalled average masses?
4. Does the trendline you calculated (D2) and drew on the graph (D6) indicate whether the masses of pennies has increased, decreased, or remained relatively the same over time?
5. If the masses of pennies had remained relatively constant over time, or varied randomly, the slope of the trendline would be 0. Is "0" in the 95% confidence interval you calculated for the slope of the trendline (D4)? What does this tell you about the general trend in penny masses over time?

6. Print out two copies of the portion of the spreadsheet containing the acceptable data and the calculations done immediately under these. *Do NOT print out the list of 300+ masses in one column*. Format your print area so that all values are printed on a single sheet of paper. Attach one copy in your notebook and hand the other in with your report.
7. Print out two copies of the graph you made. Attach one copy in your notebook and hand the other in with your report.

Homework Problems

The following problems from your book must be completed in your lab notebook (see the Syllabus for other suggested problems): Ch 4: **1**, **3**, **5**, **9**, **13**

