

## Schmidt Hammer data analysis

Last week we collected some data with the Schmidt Hammer. This week we will conduct an analysis of the data. The questions we want to answer are:

- Is there is significant statistical difference between data collected horizontally and vertically?
- Is there is significant statistical difference between sidewalk and rock number 2?
- What is the correlation between the vertical and horizontal tests?
- What is the minimum number of samples we need to take to obtain a mean significantly different (95%) from the others?

The write-up should have an introduction, method, and conclusions. The introduction should contain a brief description of the dataset (what it is, how it was collected) and the questions to be addressed. The method should describe the statistical method and assumptions. Finally, the results should have the answers to the questions posed in the introduction as well as a discussion of why (or why not) they should be believed. I encourage everyone to look up resources on the internet as well. All requested tables, etc, should be in the write-up. We want to know whether the data differs significantly for the datasets. We will do this by comparing the means,  $\mu_1$  and  $\mu_2$ . Our *hypothesis* is that  $\mu_1 \neq \mu_2$ . The opposite hypothesis (or null hypotheses) is that  $\mu_1 = \mu_2$ .

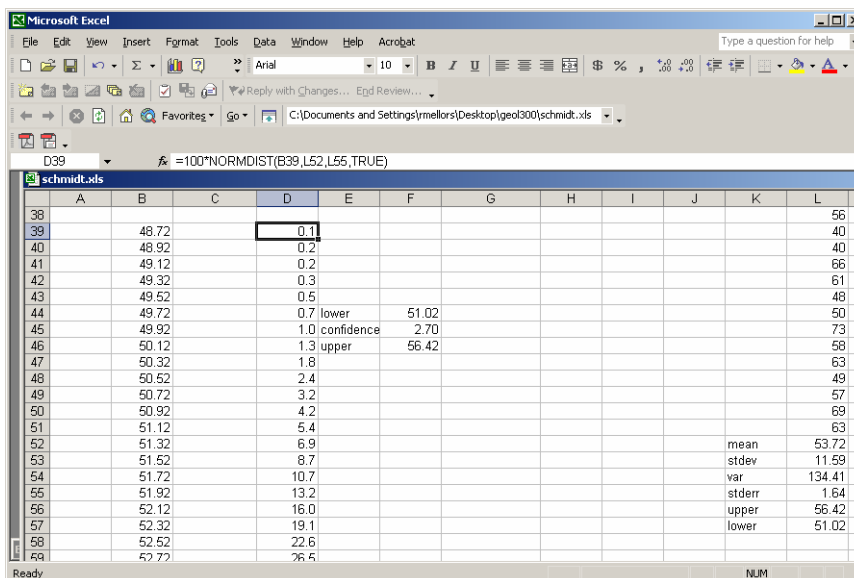
A) First, plot the data as a x,y plot. It is always good to be familiar with the data. Then calculate and show the equations for the sample average ( $\mu_s$ ), variance ( $s$ ), standard deviation ( $\sigma_s$ ), and the standard error of the data ( $s_e$ ). Calculate  $\mu_s$ ,  $\sigma_s$ , and  $s_e$  for the samples. Create a table summarizing the mean, variance, standard deviation, standard error, and number of observations for the dataset (vertical, horizontal, concrete, and other rock). Now we need to determine if the mean value of the data collected horizontally versus vertically differs at a statistically significant level. To do this we need to determine the distribution of the means.

Confidence intervals for normal distribution with known mean and standard deviation. Look at the mean value and standard deviation for the vertical and horizontal datasets (of the first rock). To find out whether the mean values are significantly different we need to determine a distribution for the mean ( $\mu$ ) itself. We will assume a normal distribution for  $\mu$ . To determine the probability we need the mean and standard deviation of the sample mean. For a large number of samples, it turns out that we can use the sample mean (from the data) for the mean of this new distribution and that the standard error (of the data) is the standard deviation of the new distribution for  $\mu$ . For a small number of samples we cannot assume that is true and we must use something called Student's T distribution. The cutoff number is generally assumed to be around 20 – 30 and so for our dataset it is not clear which we should use so we will use both.

First, we will assume that we have enough samples and we will create a plot of a cumulative normal distribution using the mean and standard error of the data. First, create a column of numbers that range from  $\mu - \sigma$  to  $\mu + \sigma$  (using your values for the mean and standard deviation for the large number of observations of the vertical rock test). This will be our **x**. Then use the **normdist(x,mean,stdev,TRUE)** command to generate a cumulative normal distribution and multiply by 100 to get percent. Plot it. For **mean** we use the mean of the combined data. For **stdev** use the standard error of the combined data (we are plotting the distribution of the mean, not the data). Do this for all four datasets (vertical, horizontal, concrete, and rock).

At what value of x is the probability 5%? What value of x is the probability 95%? (it is likely that it will not be exactly 5% at a given cell but get as close as possible with a reasonable spacing on x). Mark it on the plot. Do the means from the individual groups for vertical and horizontal fall within the range of 5% to 95% probability? If they fall outside these bounds, we could conclude that the mean values derived from using the Schmidt Hammer vertically or horizontally are statistically distinguishable with a sample value of 25 observations. The second (and much easier and more exact) method is to use the built in **confidence(probability, stdev, n)** command. Use 0.05 for the probability together with the standard deviation (of the data) and it should yield similar numbers to the first estimate (when subtracted or added to the mean). Use the **confidence()** command to generate confidence interval for the horizontal and vertical tests. Finally, compare the results with the original data – does it seem reasonable? Remember, the intervals are for the mean of the data, not the data itself.

**Figure 1.** Part of my Excel sheet to plot and calculate the bounds of the 95% confidence interval for some of the data – do not use these numbers).



B) Now assume that we cannot use the data to derive the distribution of the means (i.e. we do not know the standard deviation). In this case we use Student's T distribution to calculate. The true mean for each is expected to fall between  $\mu_s - T(df, \alpha/2)(s_{est})$  and  $\mu_s + T(DF, \alpha/2)(s_{est})$  where DF is the degrees of freedom (n-1),  $\alpha$  is the (1 - size of the confidence interval), or (1.0 - 0.95) in this case, and  $s_{est}$  is the standard error.

Excel has three functions for various use of the Student's T test: ***tdist()***, ***tinvt()***, and ***ttest()***. For this case, where we know the desired probability and degrees of freedom (number of observations - 1), we can use ***tinvt(probability, degrees\_of\_freedom)***. For example, if n=31,  $\alpha=0.1$  (90% confidence), and the sample mean is 58.61 and the sample variance is 33.43, the 90% confidence intervals are:

$$(58.61 - (TINV(0.1, 30) / \text{SQRT}(31)) * \text{SQRT}(33.43)) = 56.85 \text{ and } 60.37.$$

Calculate the confidence levels using ***tinvt()*** and the above formulas for all four cases in a table or graphically. ***tinvt()*** yields the appropriate value for  $T[(n-1), (\alpha/2)]$  so the values for  $\mu_s$ ,  $s$ , and  $n$  will be needed. See Figure 2 for an example. The confidence intervals using Student's t test should be larger than the ones calculated in part A in general.

Do the confidence ranges overlap? If they do, then we cannot distinguish different rock types (or orientations) with the Schmidt Hammer with this number of samples. Another useful function is ***ttest()***. Read the help on ***ttest()*** and apply to these two data sets. Assume unequal variance and two-tailed distribution. What answer do you get and is it consistent with your earlier results?

**Figure 2.** Confidence intervals are the mean plus or minus a value. This value can be calculated using confidence() or with a Student T test.

	A	B	C	D	E	F	G	H
27	mean	51.20	40.44	36.16	31.32		55.68	44.20
28	stdev	11.30	11.16	3.57	4.34		11.46	11.41
29	variance	127.58	124.51	12.72	18.81		131.23	130.17
30	stderr	2.26	2.23	0.71	0.87			
31								
32	stdev							
33	variance							
34	Tinv	2.0639	2.0639	2.0639	2.0639		4.4277	
35								
36	lower bound	46.54	36.83	34.69	29.53		46.77	
37	upper bound	55.85	45.05	37.63	33.11		55.63	
38								
39		48.72		0.1				
40		48.92		0.2				
41		49.12		0.2				
42		49.32		0.3				
43		49.52		0.5				
44		49.72		0.7	lower	51.02		
45		49.92		1.0	confidence	2.70		
46		50.12		1.3	upper	56.42		
47		50.32		1.6				
48		50.52		2.4				
49		50.72		3.2				
50		50.92		4.2				

C) Correlation and regression. Using the **correl()** function, determine the correlation between the vertical and horizontal rock tests. Then plot both using a scatter plot and add a linear trendline between the two series (create graph, click on data point, and add trendline). Then use the regression tool (tools, data analysis) [if data analysis does not appear, click on add-ins and add the data analysis tool pack]. Follow the directions to calculate the regression for the two data sets. What is the standard error, slope, and y-intercept of the regression line?

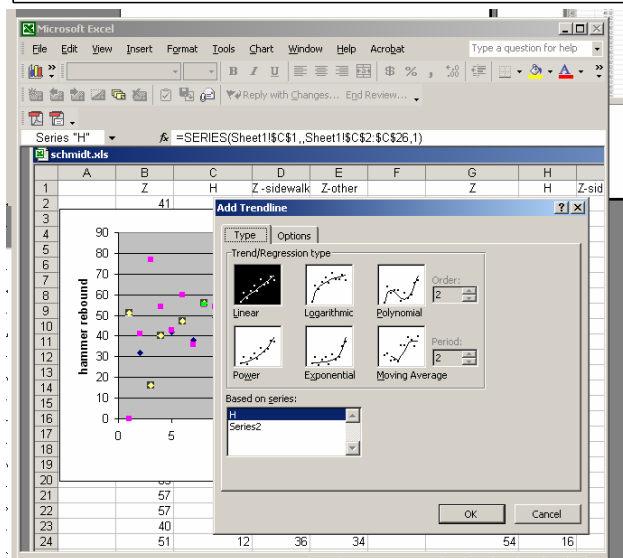
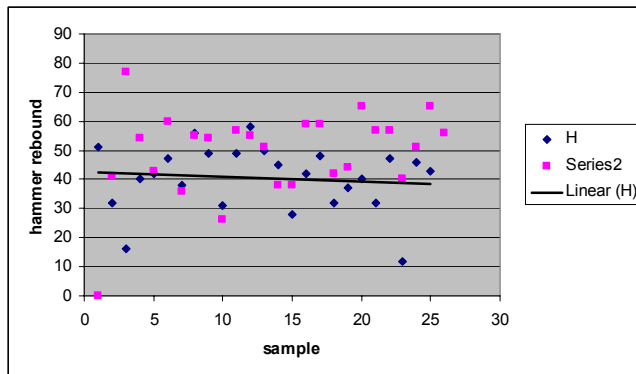


Figure 3. Adding a trend line to a chart.