Geomorphology and Aerial Photo Interpretation
Introduction to Data Analysis Using Hillslope Gradient

c:\wou\geomorph\dataex.wpd

This exercise is designed to strengthen your computer skills and provide an introduction to statistical analysis. Read over the introduction to statistical analysis, then you will be ready to work through the tutorial below.

INTRODUCTION TO DATA ANALYSIS TUTORIAL

To complete this tutorial, you will need access to a personal computer (either at home or at WOU) that has internet access, web-browser software (internet explorer or netscape), and Microsoft Excel.

A. Importing Data

1. Log onto your favorite computer at home or WOU, access your internet or network connection.

    A. To use the new Natural Science Computer Lab in RM 216...
        -at Novell Login window, click advanced and choose
            context = users.student.acad.wosc
            server = st1
        - type in username and password, click OK to log onto student server st1

NOTE: the NS 216 lab may be accessed during the day, see a faculty member. There are also evening hours from 5-7 pm with a student helper, Mon - Thursday. If no one is in the lab by 6:00 PM, the student is permitted to lock the lab and leave.

2. Use Internet Explorer and go to http://www.wou.edu/taylor Follow the links to the G322 Geomorphology class page.
3. Click on the "Data Analysis Tutorial" link. Excel should automatically open the file "example.xls"
4. Look at the data set. It consists of hillslope gradient data from two different field areas, each underlain by siltstone and sandstone bedrock, respectively.
5. Save the data as an Excel "workbook file" to the network or or floppy disk. To save on floppy:

    File - Save As-Workbook File - Sloptest.xls" on the a:\ drive

or to save on your st1 student account

    File-Save As-Workbook File - "Sloptest.xls" on the H:Eon.Stu.Folder - folder = login name

6. Close Internet Explorer
7. Start MS Excel, and open the workbook file you just saved
    -file-open-sloptest.xls (file is either on floppy on on your student account folder somewhere by now)

You should now have a data set imported into microsoft excel in 3 columns ("Bin Range", "siltstone_slopes", and "Sandstone_slopes")

B. Basic Statistical Analysis of Data

1. Use the Excel statistical functions to fill in the statistical summary of the data at the bottom of the spreadsheet. The following is a summary of Excel stat function commands (Type these commands in the appropriate cell):

*note: *cell range* is the range of cell addresses to include in the command, for example if you wanted to add all cells in column B from row 2-20; the formula is =sum(b2.b20)

Stat	Command
Mean	=average(*cell range*)
Median	=median(*cell range*)
Standard Deviation	=stdev(*cell range*)
Maximum Value	=max(*cell range*)
Minimum Value	=min(*cell range*)
No. of Observations	=count(*cell range*)

Try printing out your results:

-highlight all cells that you want to include in the print out
-file-print area-set print area
-file-print-print preview
-select printer and print

2. Frequency Analysis

Critical question: are the slopes in the tutorial exercise significantly different from one another?

Null Hypothesis = the siltstone and sandstone slopes are NOT significantly different from one another.

Let's first perform a frequency analysis using Excel. A frequency analysis measures the distribution of data across a number of ranges (or bins). Here's how to complete a frequency analysis with Excel. Double-click on an Excel icon. Follow the point-and-click instructions below:

-tools-addins-analysis tool pak (note: skip this procedure if "data analysis" is already an option under tools)

-tools-data analysis-histogram

input range: click in box, then highlight the "siltstone" column (don't include the column title)

bin range: click in box, then highlight the "bin range column" (don't include the column title)

click radio button on new worksheet ply (this will put answer on new worksheet), entitle it "histogram" in the worksheet ply box

click OK to conduct the analysis

The results for the siltstone should look like this:

Bin	*Freq*
30	0
30.5	0
31	1
31.5	0
32	2
32.5	2
33	3
33.5	2
34	0
34.5	0
35	0
More	0

The bin intervals simply refer to the base value of the frequency range, for example:

30 = no. of values between 30-30.5
30.5 = no. of values between 30.5-31
31 = no. of values between 31-31.5.... etc.

3. Now let's graph the histogram data

-highlight both the bin and freq. columns of the histogram output (include the column titles)
- click on the chart icon on the Excel tool bar
-double-click on the "clustered column" chart option
-click the radio button with "series in columns"
- next
- type in a chart title, x and y axis title (e.g. x = siltstone slopes, y = frequency)
-next
-click radio button: object in "histogram" or in the present worksheet
-finish
-click on the graph so that the box is highlighted, then pull the edges of the graph to resize it, resize the graph so that it is readable.

- now highlight the histogram data cells and the graph cells

-file-printarea-set printarea
-file-print-print preview-print
... send the work to a printer for hardcopy output.

Go back to the original slope data worksheet and repeat the histogram / frequency procedure for the sandstone slopes.

C. T-Test Analysis

A t-test is a statistical method for determining if there is a statistical difference between two sample population means (for example are the siltstone slopes significantly steeper than the sandstone slopes?). As stated above, our "null hypothesis" in this case is that there is NO significant difference between the two data groups. Here's how we will test it using Excel:

- tools - data analysis
-choose "t-test: two-sample assuming equal variances"
-Input / Variable Range 1: click in the box, then highlight the "silstone" column (include the column title)
-Input/ Variable Range 2: click in the box, then highlight the "sandstone" column (include the col. title)
- Hypothesized mean difference = 0 (i.e. we're hypothesizing no differences between means)
-click on the "labels" box, telling excel that the first row has a column title in it

-set Alpha = 0.05
((1 - alpha) x 100% = the confidence with which we will make the test, e.g. alpha = 0.05, confidence = (1-0.05) x 100% = "95%" confidence that our test results are correct)

-click on the radio button for "new worksheet ply" and type "t-test" in the box... this will tell excel to put the t-test results on a new worksheet, and it will call this worksheet "t-test"

- now click on "OK" to run the test
-set the print area-print preview-print the results

The results of the test should look like this:

t-Test: Two-Sample Assuming Equal Variances

	siltstone_slope	*sandstone_slope*
Mean	32.6	32.85
Variance	0.6	2.225
Observations	10	10
Pooled Variance	1.4125	
Hypothesized Mean Difference	0	
df	18	
t Stat	-0.47036	
P(T<=t) one-tail	0.321873	
t Critical one-tail	1.734063	
P(T<=t) two-tail	0.643747	
t Critical two-tail	2.100924	

Now, we will analyze the results. Excel t-test gives the sample mean, variance, degrees of freedom (no. of obs. - 1), a calculated "t stat", and a "critical" t-stat. To determine whether we accept or reject our null hypothesis, we compare the calculated "t stat" to the "t Critical two-tail", and should be read as + or -.

For example above, t Critical two-tail = 2.10 = either +2.10 or - 2.10 (i.e. it is two-tail)

The t-critical two-tail is a statistic from a table. If the calculated "t-stat" is greater the "t-critical", then we REJECT the null hypothesis... in this case, it would be interpreted that there IS a significant difference between the silstone and sandstone slopes. If the calculated "t-stat" is less than the "t-critical", then we ACCEPT the null hypothesis... that there is no difference between siltstone and sandstone slopes.
In our result: t-stat = -0.47 t-critical = +/- 2.10

the absolute value of the t-stat is less than that of t-critical, hence we ACCEPT the null hypothesis... i.e. the siltstone and sandstone slopes are the same statistically (at least we a 95% sure, since we used an alpha of 0.05, remember).

ANALYSIS OF APPALACHIAN HILLSLOPE DATA

Now that you know how to do it, let's try analysis of some actual slope data from the Appalachians of West Virginia and Virginia. I've gathered hillslope gradient data from two watersheds: North Fork basin in Pocahontas County, WV and Little River basin in Augusta County, VA. The data is located at the class web-site, listed as "Appalachian Slope Data". Access the data using Internet Explorer, and Saveas-Excel Workbook-Mydocuments-appslope.xls. Or alternatively, save to a floppy disk on the a:\drive. Close Internet Explorer, fire-up Excel, then file-open-appslope.xls.

Using the above tutorial as a model, conduct a statistical analysis of the Appalachian data to determine if there is a statistically significant difference between hillslopes at each area. Complete the following statistical parameters for each site, using the tools you have just used in the tutorial:

mean, median, standard deviation, maximum slope, minimum slope, number of observations, frequency distribution (use the bins provided), frequency bar graph, and t-test (use an alpha = 0.05, report the mean, variance, calculated t-stat, and two-tailed t-critical value).

Package all of your data and analyses in an organized format in Excel and print your results. The final lab report should include an organized print out of all data and analyses, for both the tutorial and Appalachian data set.

Answer the following questions:
1) Does your variance equal the square of the standard deviation?

2) What is the null hypothesis of the Appalachian slope analysis?

3) Is the calculated t-stat greater than or less than the critical two-tail t-stat?

4) How confident (what is the confidence level) of your results?

5) Is there a significant difference between hillslope gradients at the North Fork and Little River areas?

6) Given that these areas are within 100 km of each other (i.e. fairly close proximity), can you hypothesize what factors might be controlling the relationships between hillslope gradients?

7) Based on your results, which area has a greater likelihood of being associated with slope failure (e.g. landslide and debris flow)... or are both areas the same? Rank each area with respect to slope-failure potential using a relative criteria of lower vs higher.

8) Which area would likely have thicker soil/regolith deposits on the hillslopes? WHY?

STATISTICAL OPERATIONS:

Statistics are used to make inferences about the population from information about a sample of the population.

An average, or mean, is a value representative of a set of sample data. Averages are called measures of central tendency since values lie near the center of a set of data arranged according to magnitude. The mean of a set of N observations $X_1, X_2, X_3, \ldots, X_N$ is the sum of all the observations divided by the number of observations or

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots X_N}{N} = \frac{\Sigma X}{N} \tag{6.1}$$

(You've done this many times to get your average mark for a course.) If the sample is unbiased, the mean of the sample ($\bar{X}$) is closer to the mean of the total population (μ) from which the sample was drawn than any other estimate of μ based on that same sample.

The degree to which the sample data vary from the mean is measured by the standard deviation. The standard deviation(s) of a set of data $X_1, X_2, X_3, \ldots, X_N$ is defined as

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} \tag{6.2}$$

or, the standard deviation of a sample is equal to the square root of the sum of the squares of *each observation minus the mean of the observations*, divided by the total number of observations.

For ease of calculation the following formula can be used:

$$s = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} \tag{6.3}$$

stated as: standard deviation of a sample is the square root of the sum of each observation squared over the number of observations minus the sum of the observations over the number of observations squared.

We may want to make inferences about the means of two samples. That is, we may want to determine if our two samples came from the same population, or if the two populations are the same. Another way of saying this is that we wish to determine if there is a significant difference in the means of the two samples. Since small samples may not be normally distributed, a special statistic t is used to test the hypothesis that two samples come from the same population, that is, that the means of the two populations are equal, or

$$\mu_1 = \mu_2$$

In doing this a *null* hypothesis is set up: the difference between the means is not significant, or that the difference between them is due to chance alone, or again, that this difference found between the means would be expected 95% of the time (if the level of significance is 0.05). All of these statements have the same meaning.

The level of significance is the maximum probability with which we would be willing to risk an error. This probability (α) should be specified before the test is carried out. If the level of significance (α) is 0.05, then we are confident that we are right 95% of the time if we accept or reject the hypothesis.

For example, the null hypothesis for testing data taken on the slope of round and angular sand grains would be:

there is no significant difference between the angle of repose formed by angular grains of sand and the angle formed by rounded grains.

To carry out a t-test significance, we calculate t:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \tag{6.4}$$

$$\text{where } \sigma_p = \sqrt{\frac{N_1 s_1^{\,2} + N_2 s_2^{\,2}}{N_1 + N_2 - 2}} \tag{6.5}$$

Where $\bar{X}_1$ is the mean of sample #1 (larger mean)

$\bar{X}_2$ is the mean of sample #2

N_1 is the number of observations in sample #1

N_2 is the number of observations in sample #2

σ_p is the estimated standard deviation of the two populations combined

s_1 is the standard deviation of sample #1

s_2 is the standard deviation of sample #2

The degrees of freedom, f, are equal to the sum of the number of observations minus 2.

$$f = N_1 + N_2 - 2$$

STATISTICAL PROCEDURE FOR t-TEST

(1) Determine the means of each sample using equation 6.1.

(2) Calculate the standard deviation of each sample using equation 6.3.

(3) Set up a null hypothesis about the means of the samples you are testing.

(4) Determine the level of significance, α.

(5) Calculate t and f, where f is the degrees of freedom and t is calculated using equations 6.4 and 6.5.

(6) Look up t in the t-table for the α and f of the samples.

(7) Compare calculated t with t from the t-table and accept or reject the null hypothesis. Accept if calculated t is within the range of ±t from the t-table, otherwise reject.

Carry out a t-test to see if your samples are different.