302 Class Assignment – Introduction to Geostatistics and Data Analysis

Part 1 Reading Assignment

Go to the class web site and download the following reading: “Introduction to Geostatistics and Data Analysis”. Read the reference; briefly define the following terms, answer the following questions.

1. Frequency Distribution

2. Using Figure 1 on p.32 of the reading, draw a generalized, hypothetical histogram showing frequency (no. of occurrences) on the y axis and quantified scores on the x axis.

3. Tally –

4. Explain the difference between a relative frequency and cumulative frequency. Sketch a generalized diagram showing how each would be graphed.

5. Sketch the difference between a histogram graph and a frequency polygon graph.

6. Define “intervals of values”, otherwise known as “bins”

7. Make a sketch of a histogram showing a normal distribution. Where do the most numerous frequency of occurrence plot in a normal distribution?

8. Looking at the example on Figure 5 (from Table 3), create a Pie Chart showing the distribution of the following geologic data for a rock sample; complete the table:

Mineral Type	Frequency	%Frequency	Portion of Pie Chart 360 circle
Plagioclase	524	_______	_______
Potassium Feldspar	334	_______	_______
Quartz	728	_______	_______
Hornblende	123	_______	_______
Augite	56	_______	_______
Mica	224	_______	_______

9. Define the following terms, write equations where necessary:

central tendency

mode

median

mean

sample

population

tail of distribution

weighted mean

10. In a perfectly normal frequency distribution of data, how will the mean, median, and mode compare to one another?

11. Define the following terms, write equations where necessary:

Dispersion of data

Data range

Variance

Standard Deviation

12. Question: at +/- 1 standard deviation away from the mean in a normal distribution, what total percentage of occurrences will fall in the central area of occurrence about the mean?

13. Define the following terms, write equations where necessary:

Null Hypothesis

Significance level

Write the equation for a t value, what is it used for?

Type I error

Type II error

Part 2 – Reading Overview of Microsoft Excel Statistical Functions

Read back over the Microsoft Excel tutorials and class notes, Excel provides some statistical functions via the function wizard, available equations, and through use of the "Tools-Data Analysis" wizard off of the main pull down menus.

Microsoft Excel provides a set of data analysis tools — called the Analysis ToolPak — that you can use to save steps when you develop complex statistical or engineering analyses. You provide the data and parameters for each analysis; the tool uses the appropriate statistical or engineering macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

Related worksheet functions Excel provides many other statistical, financial, and engineering worksheet functions. Some of the statistical functions are built-in and others become available when you install the Analysis ToolPak.

Accessing the data analysis tools The Analysis ToolPak includes the tools described below. To access these tools, click Data Analysis on the Tools menu. If the Data Analysis command is not available, you need to load the Analysis ToolPak add-in program.

Anova

The Anova analysis tools provide different types of variance analysis. The tool to use depends on the number of factors and the number of samples you have from the populations you want to test.

Anova: Single Factor This tool performs a simple analysis of variance, testing the hypothesis that means from two or more samples are equal (drawn from populations with the same mean). This technique expands on the tests for two means, such as the t-test.

Anova: Two-Factor With Replication This analysis tool performs an extension of the single-factor anova that includes more than one sample for each group of data.

Anova: Two-Factor Without Replication This analysis tool performs a two-factor anova that does not include more than one sampling per group, testing the hypothesis that means from two or more samples are equal (drawn from populations with the same mean). This technique expands on tests for two means, such as the t-test.

Correlation

The Correlation analysis tool measures the relationship between two data sets that are scaled to be independent of the unit of measurement. The population correlation calculation returns the covariance of two data sets divided by the product of their standard deviations based on the following formulas.

You can use the correlation analysis tool to determine whether two ranges of data move together — that is, whether large values of one set are associated with large values of the other (positive correlation), whether small values of one set are associated with large values of the other (negative correlation), or whether values in both sets are unrelated (correlation near zero).

Note To return the correlation coefficient for two cell ranges, use the CORREL worksheet function.

Covariance

Covariance is a measure of the relationship between two ranges of data. The Covariance analysis tool returns the average of the product of deviations of data points from their respective means, based on the following formula.

You can use the covariance tool to determine whether two ranges of data move together — that is, whether large values of one set are associated with large values of the other (positive covariance), whether small values of one set are associated with large values of the other (negative covariance), or whether values in both sets are unrelated (covariance near zero).

Note   To return the covariance for individual data point pairs, use the COVAR worksheet function.

Descriptive Statistics

The Descriptive Statistics analysis tool generates a report of univariate statistics for data in the input range, providing information about the central tendency and variability of your data.

Exponential Smoothing

The Exponential Smoothing analysis tool predicts a value based on the forecast for the prior period, adjusted for the error in that prior forecast. The tool uses the smoothing constant a, the magnitude of which determines how strongly forecasts respond to errors in the prior forecast.

Note   Values of 0.2 to 0.3 are reasonable smoothing constants. These values indicate that the current forecast should be adjusted 20 to 30 percent for error in the prior forecast. Larger constants yield a faster response but can produce erratic projections. Smaller constants can result in long lags for forecast values.

F-Test Two-Sample for Variances

The F-Test Two-Sample for Variances analysis tool performs a two-sample F-test to compare two population variances.

For example, you can use an F-test to determine whether the time scores in a swimming meet have a difference in variance for samples from two teams.

Fourier Analysis

The Fourier Analysis tool solves problems in linear systems and analyzes periodic data by using the Fast Fourier Transform (FFT) method to transform data. This tool also supports inverse transformations, in which the inverse of transformed data returns the original data.

Histogram

The Histogram analysis tool calculates individual and cumulative frequencies for a cell range of data and data bins. This tool generates data for the number of occurrences of a value in a data set.

For example, in a class of 20 students, you could determine the distribution of scores in letter-grade categories. A histogram table presents the letter-grade boundaries and the number of scores between the lowest bound and the current bound. The single most-frequent score is the mode of the data.

Moving Average

The Moving Average analysis tool projects values in the forecast period, based on the average value of the variable over a specific number of preceding periods. A moving average provides trend information that a simple average of all historical data would mask. Use this tool to forecast sales, inventory, or other trends. Each forecast value is based on the following formula.

where:

N is the number of prior periods to include in the moving average
Aj is the actual value at time j
Fj is the forecasted value at time j
Random Number Generation

The Random Number Generation analysis tool fills a range with independent random numbers drawn from one of several distributions. You can characterize subjects in a population with a probability distribution.

For example, you might use a normal distribution to characterize the population of individuals' heights, or you might use a Bernoulli distribution of two possible outcomes to characterize the population of coin-flip results.

Rank and Percentile

The Rank and Percentile analysis tool produces a table that contains the ordinal and percentage rank of each value in a data set. You can analyze the relative standing of values in a data set.

Regression

The Regression analysis tool performs linear regression analysis by using the "least squares" method to fit a line through a set of observations. You can analyze how a single dependent variable is affected by the values of one or more independent variables.

For example, you can analyze how an athlete's performance is affected by such factors as age, height, and weight. You can apportion shares in the performance measure to each of these three factors, based on a set of performance data, and then use the results to predict the performance of a new, untested athlete.

Sampling

The Sampling analysis tool creates a sample from a population by treating the input range as a population. When the population is too large to process or chart, you can use a representative sample. You can also create a sample that contains only values from a particular part of a cycle if you believe that the input data is periodic.

For example, if the input range contains quarterly sales figures, sampling with a periodic rate of four places values from the same quarter in the output range.

t-Test

The t-Test analysis tools test the means of different types of populations.

t-Test: Two-Sample Assuming Equal Variances   This analysis tool performs a two-sample student's t-test. This t-test form assumes that the means of both data sets are equal; it is referred to as a homoscedastic t-test. You can use t-tests to determine whether two sample means are equal.

t-Test: Two-Sample Assuming Unequal Variances   This analysis tool performs a two-sample student's t-test. This t-test form assumes that the variances of both ranges of data are unequal; it is referred to as a heteroscedastic t-test. You can use a t-test to determine whether two sample means are equal. Use this test when the groups under study are distinct. Use a paired test when there is one group before and after a treatment.

The following formula is used to determine the statistic value t.

The following formula is used to approximate the degrees of freedom. Because the result of the calculation is usually not an integer, use the nearest integer to obtain a critical value from the t table.

t-Test: Paired Two Sample For Means   This analysis tool and its formula perform a paired two-sample student's t-test to determine whether a sample's means are distinct. This t-test form does not assume that the variances of both populations are equal. You can use a paired test when there is a natural pairing of observations in the samples, such as when a sample group is tested twice — before and after an experiment.

Note   Among the results generated by this tool is pooled variance, an accumulated measure of the spread of data about the mean, derived from the following formula.

z-Test

The z-Test: Two Sample for Means analysis tool performs a two-sample z-test for means with known variances. This tool is used to test hypotheses about the difference between two population means.

For example, you can use this test to determine differences between the performances of two car models.

Part 3. Homework / Lab Problems.

Students are to use Excel to complete the following problems, print out all related work with your name. This will be included in your lab portfolios. NOTE: Make sure you print out all of your work and include it!

1. Below are record high temperatures (in deg F) for 50 weather stations:

112	112	109	110	111
100	108	106	116	120
127	118	112	108	113
120	117	114	109	120
134	116	115	121	116
118	118	118	113	105
105	121	117	120	110
110	114	118	119	118
106	114	122	111	112
109	105	106	102	114

a. Construct a frequency distribution for the data (ungrouped… i.e. tally the no. of occurrences for each value recorded)
b. Using excel, create a bar graph or histogram of the distribution, labeling the axes and giving the graph an appropriate title with your name, print accordingly.
c. Using excel, create a scatter-plot line graph showing a frequency polygon of the same data.
d. Using excel, calculate the following statistical parameters, print accordingly.

Range, Mean, Mode, Median, Maximum Value, Minimum Value, Standard Deviation, Variance, Total No. of Observations.

2. a. Using the data from exercise 1 above (and Excel), construct a frequency distribution with intervals (or bins) two units wide, starting with bin 100-102.
b. Create a histogram of the distribution created in part 2a.
c. Create a frequency polygon of the same data created in part 2a.
d. which interval contains the most recorded temperatures? What is the midpoint of the interval? How does it compare to the statistical results in 1d above?

3 a. Using the ungrouped (i.e. raw data) frequency distribution from 1a above, create a cumulative frequency distribution.
b. Graph the cumulative frequency distribution from 3a using Excel, with cumulative percent on y axis, and temperature on x axis.
c. Reading the graph, which recorded temperature was exceeded by 50% of the stations? (i.e. the 50th percentile)
d. Reading the graph, which recorded temperature was exceeded by 25% of the stations? (i.e. the 75th percentile)

Other questions from data in 1 above:
What percentage of temperatures exceeded 111 deg. F?
What temperature marks the 25th percentile of the data distribution?
Is the temp. data normally distributed? How do you know?

3. Using excel, given the following values of x and their associated weights (w), calculated the weighted mean of t he data:

X	14	9	3	11	10	6	17	12	5
W	1.1	0.7	0.9	1.0	1.4	1.2	0.3	1.3	0.2

4. Consider a normal distribution of data, answer the following:

a. make a freehand sketch of a normal distribution
b. what percentage of values are within 1 standard deviation of the mean?
c. What percentage of values are above 2 standard deviations of the mean?

5. Using Excel or Grapher, graph the following data points

X	4	7	1	5	8	5	2	4	3
Y	80	92	52	76	106	100	69	71	65

a. Using regression tools, determine a best-fit line through the data, plot it on your graph and print
b. What is the slope of the best-fit line?
c. What is the y-intercept.
d. What is the best-fit equation of the line?