MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Civil and Environmental Engineering

1.017 Computing and Data Analysis for Environmental Applications

Quiz 3
Thursday, December 4, 2003

Please answer all questions on a separate piece(s) of paper with your name clearly identified:

Problem 1 (30 points)

Consider the following model of temporal fluctuations in air temperature:

$$y(t) = a_1 + a_2\cos[\pi\, t/12]\cos[\pi\, t/4380] + e$$

Where $y(t)$ is temperature in degrees C, t is the number of hours from midnight Dec 31, a_1 and a_2 are unknown regression coefficients, and e is assumed to be a zero mean normally random residual error with an unknown standard deviation σ_e. The first cosine function accounts for daily (diurnal) fluctuations while the second cosine function accounts for seasonal fluctuations. Suppose that you have a file called `temp.txt` that consists of a column of daily temperature measurements taken over a 2 year period.

Write a MATLAB program, using the internal MATLAB `regress` function, to perform a regression analysis for this problem. Make sure that your program does the following:

a) Computes least-squares estimates of a_1 and a_2 from the data in `temp.txt`.
b) Plots a regression curve on the same axes as the temperature measurements.
c) Plots on the same axes as the regression curve and temperature measurements a pair of 95% confidence interval curves for the predicted temperature, over the 2 year measurement period.

The MATLAB documentation for the `regress` function is attached at the end of this exam.

Problem 2 (10 points)

Consider the regression ANOVA table presented below. Indicate whether the regression is significant at the 95% confidence level. Compute the R^2 value.

Source	SS	df	MS=SS/df	$\mathcal{F}$	p
Regression	12.4	2	6.2	5.08	0.0469
Error	15.8	13	1.22		
Total	18.2	15			

Problem 3 (20 points)

Consider the following data:

Sample	Climatic Zone	Agricultural Activity	Groundwater nitrate concentration (deviation from background) (Mg/L)
1	Tropical	Low	2
2	Sub-tropical	Low	0
3	Sub-arctic	Medium	-4
4	Sub-arctic	Low	-2
5	Tropical	High	6
6	Sub-tropical	Medium	0
7	Sub-arctic	High	-6
8	Sub-tropical	High	0
9	Tropical	Medium	4

a) Arrange this data in a matrix appropriate as an input to the MATLAB ANOVA program `anova2`. Clearly label on the table the factors (climatic zone, agricultural activity) and treatments for the ANOVA problem.

b) From an inspection of the data only indicate which factors (if either) have a significant effect on nitrate concentration. DO NOT PERFORM A DETAILED ANOVA ANALYSIS – just arrive at a conclusion by looking at the data and considering what constitutes a significant effect.

The MATLAB documentation for the `anova2` function is attached at the end of this exam.

Problem 4 (20 points)

Suppose that you conduct a survey of potholes in roads in Massachusetts vs New Hampshire. You pick several 1 mile strips in each state and count the number of potholes in each strip that could do serious damage to your car. The results are:

Massachusetts:	6 6 4 8 7
New Hampshire:	4 2 5 2 6

Suppose that these results are samples of two independent **normally distributed** random variables (number of potholes/mile in Massachusetts and New Hampshire). Use a two-sided small sample *t* test to test the hypothesis *H0* that the means of these random variables are the same (i.e. that the difference between the two means is zero). Identify and evaluate the *t* distributed standardized test statistic needed to obtain a *p* value for the test. Be sure to use the correct standard deviation when you construct the test statistic from the data. Derive the *p* value from the *t* statistic plot (based on ν = 5 – 2 = 3 degrees of freedom) given at the end of this exam.

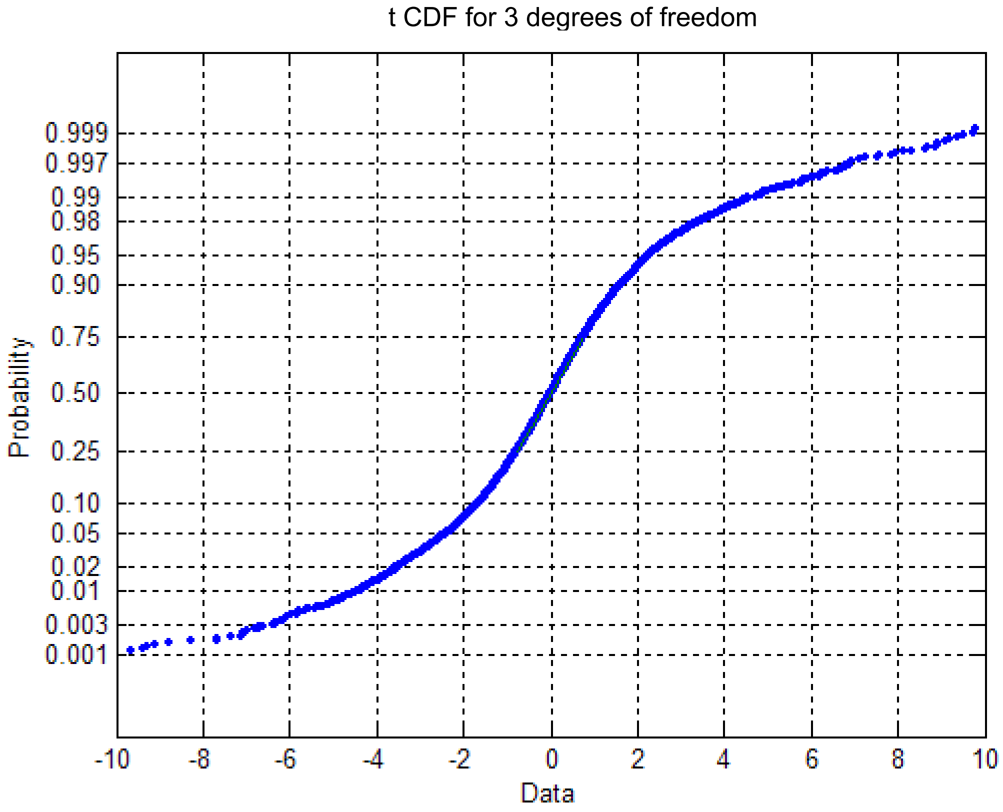
Problem 5 (20 points)

Reconsider the pothole problem described above. Suppose that you want to test the hypothesis *H0* that the mean number of potholes on Massachusetts roads is equal to the federal standard of 2 potholes per mile. Also, suppose that your Massachusetts data are samples of an **exponentially distributed** random variable (the number of potholes/mile in Massachusetts). Construct a MATLAB program to derive the CDF for your test statistic. For this problem only provide the MATLAB code needed to derive the test statistic CDF. You do not need to make any calculations.

Your program should use stochastic simulation to generate an ensemble of small sample test statistics, each constructed from a set of 5 exponentially distributed random variables. Generate these random variables with the internal MATLAB function `exprnd`. Estimate the exponential distributional parameter required by `exprnd` from the5 Massachusetts observations provided in Problem 4. Include in your program the capability to plot your derived test statistic CDF using `normplot`.

Why can't you use the *t* statistic for this problem?

The MATLAB documentation for the `exprnd` and `normplot` functions is attached at the end of this exam.



REGRESS Multiple linear regression using least squares.

b = REGRESS(y,X) returns the vector of regression coefficients, b, in the linear model y = Xb, (X is an nxp matrix, y is the nx1 vector of observations).

[B,BINT,R,RINT,STATS] = REGRESS(y,X,alpha) uses the input, ALPHA to calculate 100(1 - ALPHA) confidence intervals for B and the residual vector, R, in BINT and RINT respectively. The vector STATS contains the R-square statistic along with the F and p values for the regression.

The X matrix should include a column of ones so that the model contains a constant term. The F and p values are computed under the assumption that the model contains a constant term, and they are not correct for models without a constant. The R-square value is the ratio of the regression sum of squares to the total sum of squares.

ANOVA2 Two-way analysis of variance.

ANOVA2(X,REPS,DISPLAYOPT) performs a balanced two-way ANOVA for comparing the means of two or more columns and two or more rows of the sample in X. The data in different columns represent changes in one factor. The data in different rows represent changes in the other factor. If there is more than one observation per row-column pair, then then the argument REPS indicates the number of observations per "cell". A cell contains REPS number of rows. DISPLAYOPT can be 'on' (the default) to display the table, or 'off' to skip the display.

For example, if REPS = 3, then each cell contains 3 rows and the total number of rows must be a multiple of 3. If X has 12 rows, and REPS = 3, then the "row" factor has 4 levels (3*4 = 12). The second level of the row factor goes from rows 4 to 6.

[P,TABLE] = ANOVA2(...) returns two items. P is a vector of p-values for testing row, column, and if possible interaction effects. TABLE is a cell array containing the contents of the anova table.

To perform unbalanced two-way ANOVA, use ANOVAN.

EXPRND Random matrices from exponential distribution.

R = EXPRND(MU) returns a matrix of random numbers chosen from the exponential distribution with parameter MU.
The size of R is the size of MU.
Alternatively, R = EXPRND(MU,M,N) returns an M by N matrix.

NORMPLOT Displays a normal probability plot.

H = NORMPLOT(X) makes a normal probability plot of the data in X. For matrix, X, NORMPLOT displays a plot for each column.
H is a handle to the plotted lines.

The purpose of a normal probability plot is to graphically assess whether the data in X could come from a normal distribution. If the data are normal the plot will be linear. Other distribution types will introduce curvature in the plot.