DATA ANALYSIS IN ENGINEERING TECHNOLOGY

James E. Maisel
East Campus, Arizona State University
Mesa, AZ 85206

Abstract

A data analysis graduate/undergraduate course has been developed in the Department of Electronics and Computer Engineering Technology at the East Campus of Arizona State University. Various statistical techniques are explored to show the relevance and importance of extracting important information from raw data.

Introduction

Data analysis has permeated essentially all industrial processes. With data retrieval systems available, large amounts of data can be stored and viewed later for analysis. Raw data sets have to be processed to characterize the important data features buried in the raw data. This is where data analysis plays a key role.

Data processing is becoming a very important facet for engineering technologists. At some point in their professional career, they will be faced with using data analysis or using the results of data analysis to study the behavior of a manufacturing process [1]. In either case, their expertise in data analysis may give them the competitive edge in industry.

The Department of Electronics and Computer Engineering Technology at Arizona State University introduced a new course this past year called Data Analysis. It assumes that the engineering technologist has a rudimentary background in probability and statistics, and has senior or graduate departmental standing. A data analysis project, with a written report, differentiates the graduate from the undergraduate student.

Topics in data analysis involve tedious calculations when the data sets become large. Thus, hand calculations are restricted to very small data sets and are used to demonstrate the significance of a particular statistic.

Once the students understand the basics of, and the significance of data analysis, they are ready to use a statistical software package. As a homework assignment, they start by doing a small data set analysis using hand calculations and a software package. A comparison of results gives the students confidence in data analysis. Matlab along with Statistics Toolbox ™ was adopted because the students are familiar with Matlab from other courses taught in the department.

Table I lists the more important topics covered in the course.

Table I

Mean and variability in a data set
Scatter plots, linear correlation, curve fitting
Probability and random variables
Statistics and random sampling
Estimation using a single sample
Hypothesis testing using a single sample
Inferences using two independent samples
Simple linear regression and correlation
Inferential methods
Analysis of variance

Selected classroom topics and homework assignments will be presented in the next section. Because of the constraints on the length of this paper, the intent of these examples is to give the reader a global view of the course without getting too involved with the details of statistics.

Examples

Basics of Matlab

One of the early assignments is familiarizing the students with some of the basics of Matlab. An M-file is developed using two row vectors x and y which represent an x and y data set (Figure 1-a). Some of the main features of the assignment are: determining the mean, standard deviation, fitting a 2nd and an 8th order polynomial to the data, and generating a labeled plot of the results (Figure 1-b). Typing the M-file name asee_ex1 from the command window generates the results shown in Figure 1-a under results. The variables, p2 and p8, list the coefficients of the 2nd and 8th order polynomial, respectively, starting with the highest power first.

According to the Figure 1-b, the 2nd order polynomial is a gentle curve that follows the data trend. However, the 8th order polynomial varies more between the data points. This suggest that the 8th order polynomial is not a good function for interpolation between data points.

Fundamentals of the probability density function

One of the more important aspects of data analysis is the probability density function (pdf).In this example the students compare the difference between the normal pdf and the tpdf or sometimes referred to as the Student's pdf. The normal pdf is completely specified by the population's mean and it's standard deviation. However, if the data size is less than 30, the tpdf is used instead. The shape of the tpdf depends on the degrees of freedom, which is one less than the size of the data set.

In this assignment the students compare the difference between the two distributions. The standard normpdf (μ=0, σ=1) and a 4th degree of freedom tpdf will be used in this example. Figure 2-a is the M-file program list (asee_ex2) and Figure 2-b is the plot of both pdfs. The plot

shows that the tails of the tpdf extend farther than the normpdf and the tpdf curve is lower near zero. This must be the case since the total probability (total area) under any pdf curve must be unity.

Hypothesis testing

Hypothesis testing is a procedure for determining if an assertion about a population statistic is tenable. In this assignment the students use the Matlab Mat-file called gas.mat. Gas.mat lists the price of gas at 20 randomly chosen stations for January and February (1993) in the state of Massachusetts. Using this Mat-file, an M-file, called asee_ex3, (Figure 3-a) was written to use hypothesis testing, determine the mean and standard deviation for January and February, and whether the data for both months are normal.

The objective of this analysis is to determine if the statement of $1.15 per gallon (null hypothesis) for January or February is true. It will be assumed that the alternative hypothesis is an average price greater than $1.15. The sample mean and sample standard deviation for January and February are respectively $1.15, $0.039 and $1.18, $0.037 according to asee_ex3. Remember that the average price will vary from one 20-sample gas station price to another and it will not be exactly $1.15 due to price variability from station to station. Suppose the average price was $1.18 (February). Is the $0.03 spread a result of chance variability, or is the null hypothesis assertion correct?

In order to understand hypothesis testing fully, a significance level (typically 0.05) and the p-value must be introduced. A given significance level of 5 % means that the "probability of incorrectly rejecting the null hypothesis ($1.15), when it is actually true, is 5 %". The p-value is the probability of observing the given sample under the assumption that null hypothesis ($1.15) is true. If the "p-value < significance level", reject the null hypothesis. However if the "p-value > significance level", there is insufficient evidence to reject the null hypothesis.

The results of the hypothesis test are shown in Figure 3-a. For January the Boolean variable h =0. When h=0 the null hypothesis is not rejected. However, if h=1 the null hypothesis is rejected at the significance level of 5%. The $1.15 is within the 95% confidence interval (1.134 1.169). However, for February, h=1, which indicates that the null hypothesis can be rejected at the significance level of 5%. Note that the low confidence point (1.1675) is greater than 1.15, which puts the null hypothesis parameter outside the 95% confidence interval.

Another point that arises in data analysis is the type of distribution of the data set. The qqplot displays a quantile-quantile of two sets of samples. If the two sets have the same distribution, the plot will be linear. Assuming that the x-axis in a qqplot is a standard normal distribution, Figure 3-b shows that the data set has a very strong normal distribution.

Development of a function M-file

As part of a class discussion, the ttest2.m M-file or function file was analyzed. A function file, like ttest2.m, permits new functions to be added to the existing function file library. (The Statistic

Toolbox is a series of statistical M-files that increase the capability of the basic Matlab software.) The ttest2.m M-file is useful in hypothesis testing of a difference between two normal population means that have equal standard deviations.

The class, as an assignment, developed an M-file ttest2a.m that removes the equal population standard deviation assumption. A program listing of ttest2a.m is shown in Figure 4. Its basic output is similar to ttest2.m in Matlab (Boolean h parameter, P-value, confidence interval) with the following inputs: x (data set), y (data set), alpha (level of significance), tail (upper, lower, or two tail), and hyp (the null hypothesis value). Calculator and ttest2a.m results of a homework assignment were compared to demonstrate that ttest2a.m did perform as expected.

Conclusion

Data analysis is becoming a very important part of engineering technology. Students are given the basics of data analysis via the data analysis course necessary for their professional development.

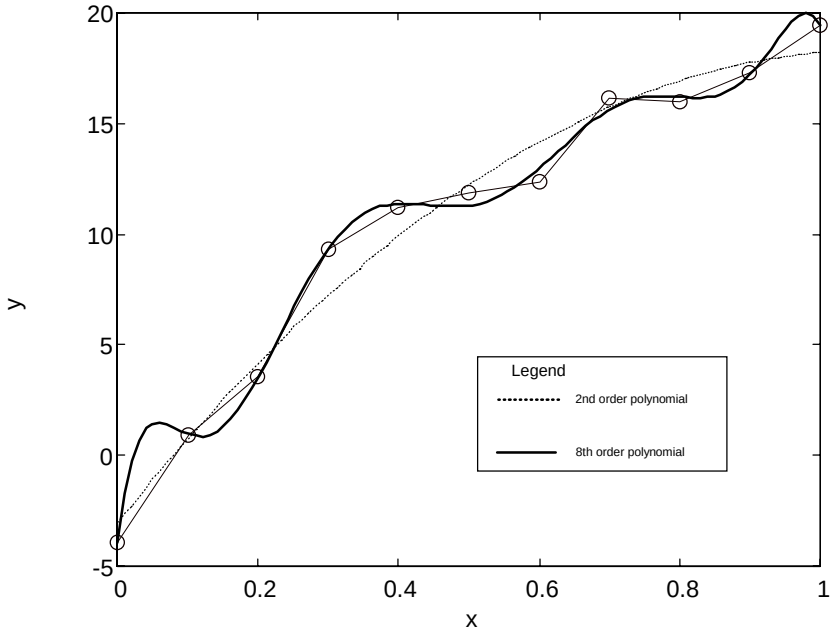
References

[1] J. Devore and R. Peck, Statistics: *The Exploration and Analysis of Data*, 2nd edition, Duxbury Press, Belmont CA., 1993, ISBN 0-534-19614-4.
[TM] Matlab and Statistics Toolbox, MathWorks, Inc, 24 Prime Park Way Natick, MA 01760-1500.

M-file: asee_ex1

```
x=[0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0]
y=[-3.89 0.956 3.56 9.32 11.16 11.82 12.32 16.12 15.96 17.3 19.4]
avgy=mean(y) % mean of y
stdev=std(y) % standard deviation of y
p2=polyfit(x,y,2) % determine the coefficients of a 2nd order polynomial
p8=polyfit(x,y,8) % determine the coefficients of an 8th order
polynomial
xi=linspace(0,1,100) % x axis data for plotting
z2=polyval(p2,xi) % evaluate 2nd order polynomial
z8=polyval(p8,xi) % evaluate 8th order polynomial
plot(x,y,'o',x,y,xi,z2,':',xi,z8) % plot information
xlabel('x'), ylabel('y') % labeling axes
gtext('Figure 1-b. 2nd and 8th order curve fitting') % mouse placement
of
                                                    text on graph
gtext('2nd order polynomial')
gtext('8th order polynomial')

RESULTS:
avgy = 10.3660

stdev = 7.3475

p2 =   -1.8589e+001   3.9842e+001   -3.0490e+000

p8 =   -3.8464e+004    1.5991e+005    -2.7143e+005    2.4128e+005
-1.1952e+005    3.2312e+004    -4.3172e+003      2.5525e+002
-3.8931e+000
```

(a)



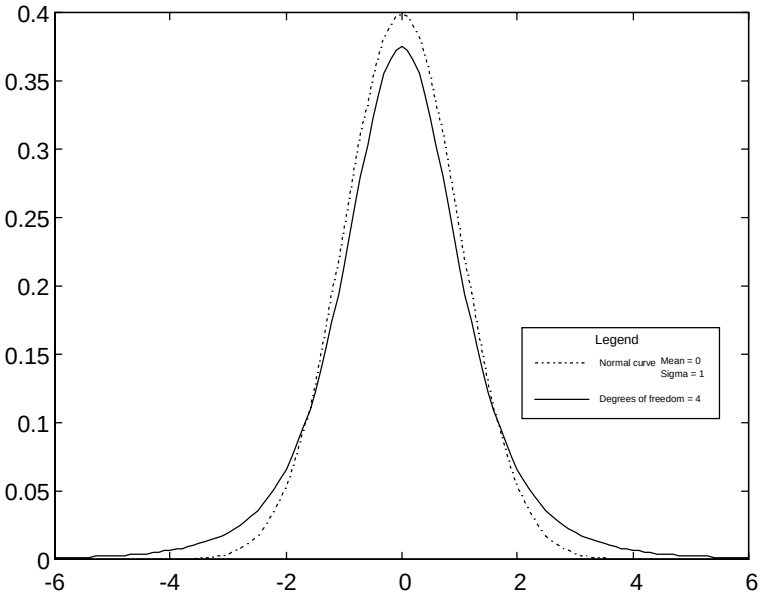
(b)

Figure 1. Program and Plot of 2nd & 8th Order Curve Fitting

M-file: asee_ex2

```
x = -6:.1:6;

y= tpdf (x,4); % Students pdf

z=normpdf (x,0,1); % normal pdf with mean =0 and standard
                   deviation = 1

plot (x,y,'-',x,z,'-.') % plotting information

gtext ('Figure 2. A comparison of the tpdf and normpdf.')
      % mouse placement on graph

gtext ('degrees of freedom = 4')
```

(a)



(b)

Figure 2. Program and Plot of the tpdf and NORM pdf

M-file: asee_ex3

```
load gas
prices=[price1 price2]
janprice = price1./100 % January prices of gas at 20 randomly chosen
 stations
febprice = price2./100 % February prices of gas at 20 randomly chosen
 stations
janavg=mean(janprice) % Average price for January
janstd=std(janprice) % Price standard deviation for January
febavg=mean(febprice) % Average price for February
febstd=std(febprice) % Price standard deviation for February
[h,pvalue,ci]=ztest(janprice,1.15,0.04) % Hypothesis testing for the
 mean of one sample with known variance
[h,pvalue,ci]=ztest(febprice,1.15,0.04) % Hypothesis testing for the
 mean of one sample with known variance
r=normrnd(0,1,20,1); % Random numbers from a normal distribution
qqplot(r,janprice); % Quantile-quantile plot of janprice vs. random
 numbers from a normal distribution
qqplot(r,febprice); % Quantile-quantile plot of febprice vs. random
 numbers from a normal distribution
```

RESULTS:

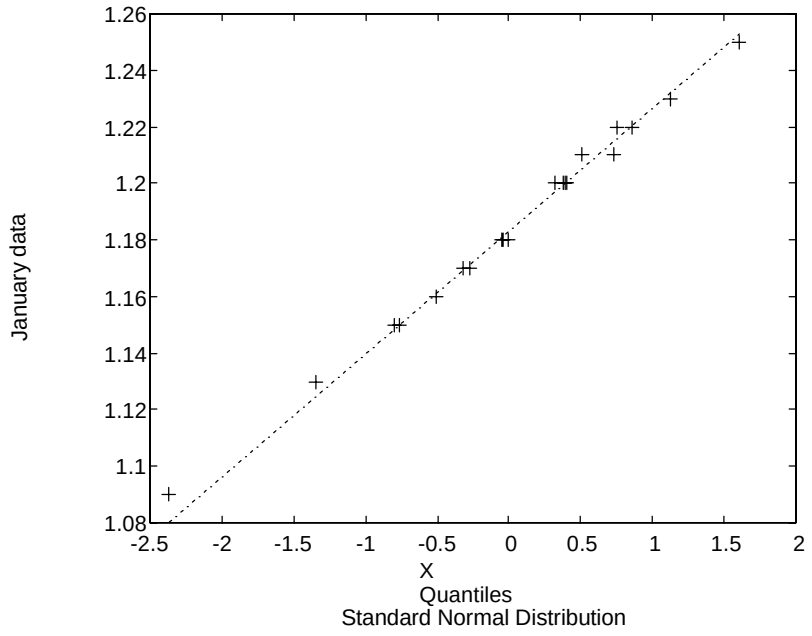
January

```
h = 0
pvalue = 0.8668
ci = 1.1340   1.1690
```

February

```
h =1
pvalue = 9.112E-005
ci = 1.1675   1.2025
```

(a)



(b)

Figure 3. Program and Quantile-quantile Plot

M-file: ttest2a

```
function [h, significance, ci] = ttest2a(x,y,alpha,tail,hyp)
%TTEST2a Hypothesis test: Compares the averages of two samples when the population variances are not equal, this program will
% also account for the modified degrees of freedom to set the  rejection region for H0: average1 - average2 = hyp  hyp = the
% hypothesized value
%   [H,SIGNIFICANCE CI] = TTEST2a(X,Y,ALPHA,TAIL) performs a t-test  to determine whether two samples from a normal distribution
%  (with unknown and different variances) could have the same mean.  The null hypothesis is: "means are equal to hyp value".
%   hyp = 0 by default  For TAIL =  0  the alternative hypothesis is: "means are not  equal to the hypothesized value."
%   For TAIL =  1, alternative: "mean of X minus the mean of Y is greater than the hypothesized value."
%   For TAIL = -1, alternative: "mean of X minus the mean of Y is less than the hypothesized value."
%   TAIL = 0 by default.   ALPHA is desired significance level (ALPHA = 0.05 by default).
%   SIGNIFICANCE is the probability of observing the given result by chance given that the null hypothesis is true. Small values of
%   SIGNIFICANCE cast doubt on the validity of the null hypothesis.
%   H=0 => "Do not reject null hypothesis at significance level of alpha."
%   H=1 => "Reject null hypothesis at significance level of alpha."
if nargin < 2,
   error('Requires at least two input arguments');
end
[m1 n1] = size(x);
[m2 n2] = size(y);
if (m1 ~= 1 & n1 ~= 1) | (m2 ~= 1 & n2 ~= 1)
   error('Requires vector first and second inputs.');
end
if nargin < 5, % this defaults the hypothesized value to 0
   hyp = 0;
end
if nargin < 4,  % this defaults the tail to a two tailed test
   tail = 0;
end
if nargin < 3,   % this defaults alpha to .05
   alpha = 0.05;
end
```

```
if (alpha <= 0 | alpha >= 1)
   fprintf('Warning: significance level must be between 0 and 1\n');
   h = NaN;
   sig = NaN;
   return;
end
c = var(x)/length(x)/(var(x)/length(x) + var(y)/length(y)); % this calculates the constant C used to compute the degrees of freedom.
dfx = length(x) - 1;
dfy = length(y) - 1;
dfe  = dfx * dfy/(dfy*c^2+(1-c)^2*dfx); % this calculates the modified degrees of freedom value
dfe = fix(dfe) % this rounds down the dfe to the nearest integer
difference = mean(x) - mean(y) - hyp;
num = sqrt(var(x)/length(x)+var(y)/length(y));
teststat   = difference / num % this calculates the test statistic for unequal variances
criticalvalue = tinv(1 - alpha/2,dfe); % criticalvalue determined by alpha and modified dfe
spread = criticalvalue * num;
ci = [(difference - spread) (difference + spread)];
significance  = 1 - tcdf(teststat,dfe);
% Adjust the significance probability for other null hypotheses.
if tail == -1
   significance = 1 - significance;
elseif tail == 0
   significance = 2 * min(significance,1 - significance);
end
% Determine if the actual significance exceeds the desired significance
h = 0;
if significance <= alpha,
   h = 1;
end
if isnan(significance),
   h = NaN;
end
```

Figure 4. Program for Hypothesis Test with different Variances

James E. Maisel, Professor, Department of Electronics and Computer Engineering Technology, College of Technology and Applied Sciences, East Campus, Arizona State University, 6001 S. Power Rd., Mesa, Arizona 85206, maisel@asu.edu.

JAMES E. MAISEL

James E. Maisel received the B.E.E. and B.E.S. degrees from Cleveland State University, in 1955, and M.S.E.E. degree from Ohio State University, in 1957. From 1958 to 1985 he was a professor of Electrical Engineering and now Professor Emeritus at Cleveland State University. He is a member of Tau Alpha Pi (faculty advisor), and is a registered Professional Engineer in the states of AZ and OH. He has the grade of Senior Member in IEEE and is a member of the ASEE.