TEACHING STATISTIC USING MATLAB

Oni Yuliani
Study Program of Electrical Engineering, Institut Teknologi Nasional Yogyakarta
Jalan Babarsari Caturtunggal, Depok, Sleman, Yogyakarta 55281
onisudaryono@gmail.com

Abstrak

Artikel ini menggambarkan penggunaan Matlab sebagai media untuk pengajaran statistika pada pendidikan sarjana. Aplikasi perangkat lunak dalam pengajaran topik-topik yang sulit seperti konsep probabilitas, distribusi probabilitas, signifikansi, dan uji signifikansi, dapat ditunjukkan dengan Matlab. Matlab telah membuktikan dirinya sebagai salah satu media yang efektif dalam proses pengajaran karena menawarkan media yang sederhana namun sangat baik dalam menganalisis dan menampilkan hasil simulasi numerik dan pengukuran

Kata Kunci: Pengajaran statistika, Matlab, proses pengajaran.

Abstract

This paper is to describe the use of Matlab as a scientific tool for the teaching of statistics in undergraduate school. Application of this software in the teaching of some difficult topics like probability concepts, probability distribution, statistical significance, and significance tests, were demonstrated using the Matlab. Matlab has proved itself to be a very effective tool in the educational process because it offers a simple and powerful tool for analyzing and visualizing results of numerical simulations and measurements.

Keywords: Teaching statistic, Matlab, educational process.

1. Introduction

In the information era, there is a thoughtful need for statistically educated students. Conventional statistical education has focused on evolving knowledge and on methodological skills, procedures, and computations. It was supposed that students would add value to the subject in the process of learning. But this approach has not functioned and does not lead students to reason or think statistically [1].

Over the past few decades there has been increasing thought given to the teaching aspects of statistics education [2]. It is broadly recognised that statistics is one of the most important quantitative subjects in a undergraduate curriculum [3]. It is also recognized that teaching statistical courses is puzzling because it serve students with variable backgrounds and abilities, many of whom have had negative experiences with statistics.

Statistics should be taught as a laboratory science explored thorough a model for developing case labs for the use of the undergraduate statistics class [4]. In this approach, it were exposed the concepts of statistics and probability through case labs. At first it is taught the theory in a lecture format and then performed the relevant lab. This will help to learn all aspects and extensions of the statistics and be prepared with the tools needed for the lab portion of the class. In this lab based lecture class it is learned the fundamental ideas of statistics in the context of contemporary real world situations.

Technology tools are increasingly becoming available to enhance and promote statistical understanding but most of them have one in common today – computer [5]. Computer based learning has found a way in the learning process in undergraduate school [6]. Computers are also significantly involved in teaching technology serving sciences like statistics. Many software applications are accessible for teaching such as Matlab. Matlab is considered as standard in technical computing and science.

Matlab is a very powerful software package that has many built in tools for solving problems and developing graphical illustrations [7]. It is also noted that the simplest method for using the Matlab product is interactively; an expression is entered by the user and Matlab immediately responds with a result. Moreover, it is possible to write scripts and programs in Matlab, which are essentially groups of commands that are executed sequentially.

The aim of the study is to demonstrate different ways of applying Matlab software in the teaching of statistics. Specifically, the study intends to demonstrate how to test hypothesis and to analyse error, to the the significanes test, and to find the correlation between two or more variables.

This paper is organised as follows: Section 2 describes a research method. Research and analysis is presented on Section 3, while Section 4 gave the conclusion followed by the reference.

Received August 5, 2019; Revised December 11, 2019; Accepted December 11, 2019

2. Research Method

A Matlab is a high-performation language for technical computing. Matlab stands for *matrix laboratory*, which reflects its original application to matrix applications [8]. From the start menu, when Matlab software is clicked on, a window opens in which the main part is the command window (Figure 1). In the command window, one should see: >> which is called prompt. In the command window, Matlab can be used interactively. At the prompt, any Matlab command or expression can be entered, and Matlab will immediately respond with the result. During this process, some commands can serve as will introduction to MATLAB and allow one get help: *Info* will display contact information for the product, **Demo** has demos of some of the features of MATLAB, *Help* explain any command; help help will explain how help works, and then *Helpbrowser* opens a help window, and finally *Lookfor* searches through the help for a specific word or phrase (note: this can take a long time). To get out of Matlab, either type quit at the prompt, or choose file, then Exit Matlab from the menu. While Figure 2 shows the command window with some basic mathematical tasks performed. Figure 3 shows an empty script file. Scripts in Matlab are used to write basic code to implement some mathematical tasks so it can be saved and can also be edited.

Vectors and matrices are used to store sets of values, which are the same type. A vector can be either a row vector or a column vector. A matrix can be pictured as a table of values. The dimensions of a matrix are *r x c*, where *r* is the number of rows and *c* is the number of columns. This is marked "*r by c*", if a

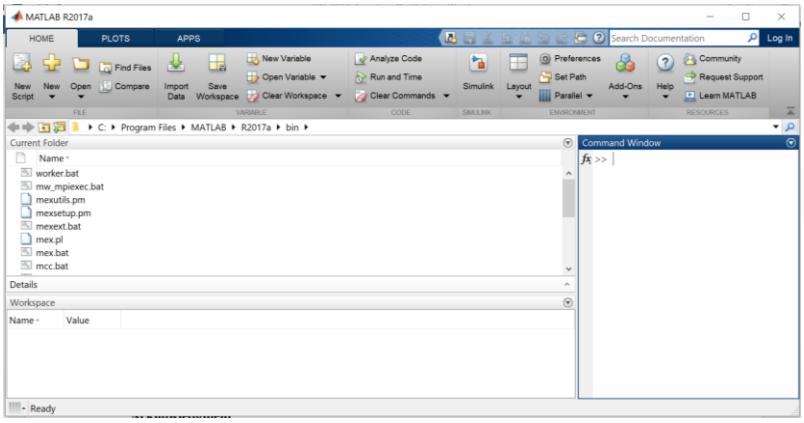


Figure 1. An empty Matlab command window

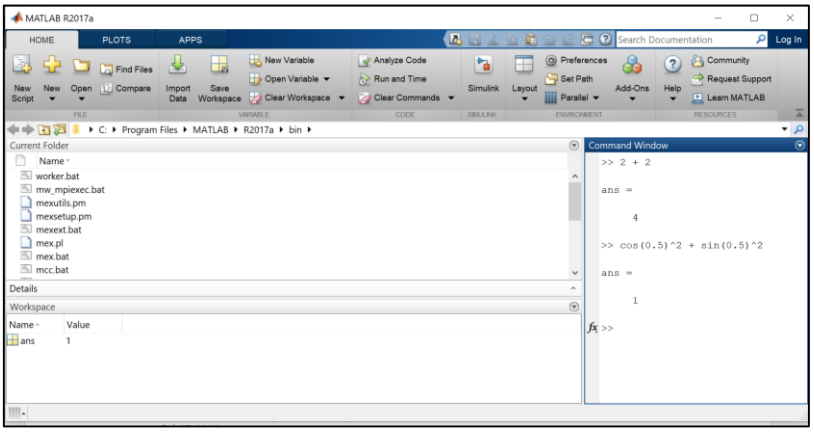


Figure 2. A Matlab command window with some commands given

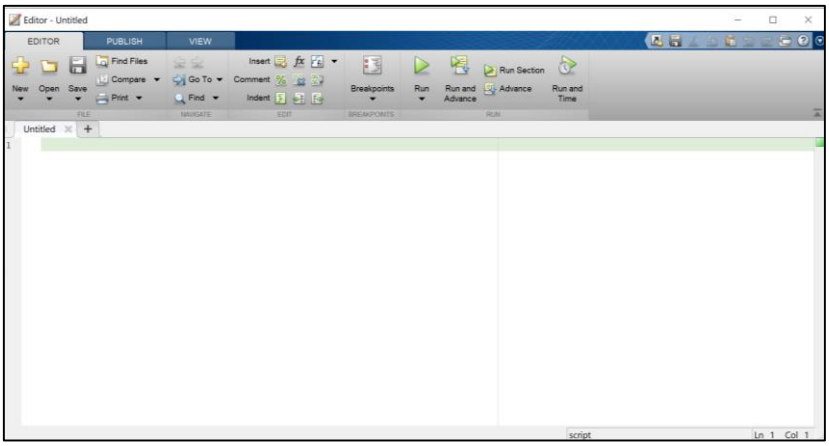


Figure 3. An empty script

vector has n elements, a row vector would have the dimensions *1 x n*, and a column vector would have the dimension *n x 1*. Matlab is written to work with matrices; the name Matlab is a short form of "matrix laboratory." Since Matlab is written to work with matrices, it is very easy to create vector and matrix variables, and there are many operations and functions that can be used on vectors and matrices.

Turning engineering data into knowledge requires an ability to test *hypotheses* and to *analyse errors*, which requires some understanding of probability and certain basic statistics. An underlying concept in statistics is that of a random variable. *Random variables* may be thought of as physical quantities, which are yet to be known. Since their values cannot predict, one may say that they depend on "chance".

After collecting a series of data, these data are regarded as a *set* in probability theory, defined as a collection of objects about which it is possible to determine whether any particular object is a member of the set. In particular, the possible result of a series of measurements (or experiments) represent a set of *points* called the *sample space*. These points may be grouped together in various ways called *events*, and under suitable conditions *probability functions* may be assigned to each. The probabilities always lie between zero and one, such that an *impossible event has the probability of zero*, and the *probability of a certain event is one*.

When a sample space of points is considered to represent the possible outcomes of a particular series of measurements, a *random variable x(j)* is a set function defined for points *k* from the sample space. A random variable *X* can assume *x(j)* values which can be real numbers between -8 and +8 , associated to each sample points that might occur. In other words, the random outcome of an experiment, indexed by *j*, can be represented by a discrete distribution of real numbers *x(j)*, which are the possible values of *X*. A random variable is described by a function called the *probability density function* (PDF). The PDF is a measure of the density of probability of the random variable plotted on a horizontal axis, which is the domain of possible values of the random variable. Thus if *X* is a random variable, the PDF *f(x)* has a graph whose area is 1, since it is certain that *x* will have some value within its domain.

Each *x(j)* has a probability *p(j)*. The discrete distribution function *f(x)* of *p(j)* is:

$$f(x) = p_j \qquad \text{if } x = x_j \qquad (j = 1,2, \dots 3) \tag{1}$$

So the probability distribution function by taking sums:

$$f(x) = \sum_{x_j s_x} f(x_j) = \sum_{x_j s_x} p_j \tag{2}$$

The first moment of a probability distribution is the *mean*, and the first central moment about the mean is zero. The second moment about the mean is called the *variance*, σ_x^2, and its square root, *sx*, is called the *standard variation* , σ_x. The third moment about the mean is called *skewness*, γ, and is zero for PDF's which are symmetric about the mean. The fourth moment about the mean is the *kurtosis*. It measures "peakedness" of the distribution.

First central moment: $$E(x - \mu) = 0 \tag{3}$$

Second central moment: $$\sigma_x = E(x^2) - \mu^2 \tag{4}$$

Third central moment: $$\gamma = \frac{1}{\sigma^3} E([x - \mu]^3) \tag{5}$$

The median value divides the probability density distribution in two halves such that there is a 50% chance for x to be less than the median and a 50% chance for it to be greater than *mx*. The Figure 4 shows measures of central tendency of an arbitrary probability density distribution.

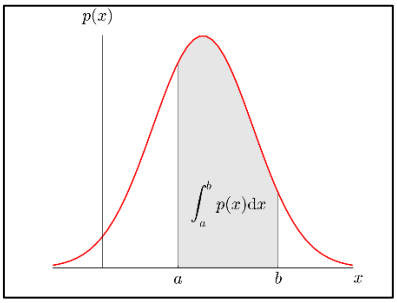


Figure 4. An arbitary probability density distribution [9]

The statistical significance (*p*-level) of a result is an estimated measure of the degree to which it is *true*, in the sense of representative of the population. The value of the *p*-level represents a decreasing index of the reliability of a result. The higher the *p*-level, the less one can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population. Definitely, the *p*-level represents the probability of error that is involved in accepting our observed result as valid as representative of the population.

There is no way to circumvent arbitrariness in the final decision as to what level of significance. The selection of some level of significance is arbitrary. In practice, the final decision usually depends on whether the outcome was predicted a priori. Typically, in many sciences, results that yield $p = 0.05$ are considered borderline statistically significant but remember that this level of significance still involves a pretty high probability of error (5%). Results that are significant at the $p = 0.01$ level are commonly considered statistically significant, and $p = 0.05$ or $p = 0.001$ levels are often called highly significant.

The Gaussian or Normal distribution is important because many natural processes result in data that are normally or log-normally distributed. The distribution of many test statistics is normal or follows some form that can be derived from the normal distribution. A random variable is said to follow a Gaussian (or normal) distribution, if its probability density function is given by [9]:

$$P(x) = (2\pi\sigma)e^{-0.5(x-\mu_x)^2/\sigma^2} \quad (6)$$

The Gaussian PDF is completely specified by the mean *mx* and standard deviation *s*. The shape of the Gaussian distribution is a bell-shaped curve, symmetric about the mean, with 68% of its area within one standard deviation, and 95% within two standard deviations, shown in Figure 5.

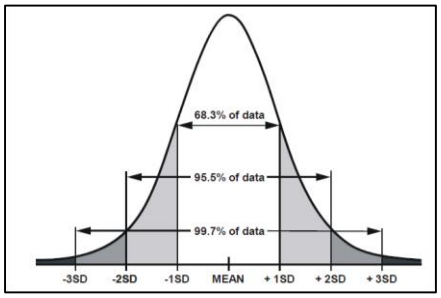


Figure 5. A Normal distribution [9]

In a Normal distribution, observations that have a standardized value of less than -2 or more than +2 have a relative frequency of 5% or less. Standardized value means that a value is expressed in terms of its difference from the mean, divided by the standard deviation.

The Gaussian distribution is quite frequently used in engineering, because of a result known as the *central limit theorem*, which states that the sum of many independent random variables tends to behave as a Gaussian random variable. This result implies that *any physical process which is the sum of random events is Gaussian in its distribution*. Inappropriately this assumption often does not embrace for some distributions of real data which have to cope with.

Most computers contain a random number generator, which produces numbers between 0 and 1 with an approximately uniform distribution, shown in Figure 6. Random numbers are approximately Gaussian with zero mean and unit variance. With a computer program to generate uniformly distributed *random numbers on the interval (0,1), one may compute the sum of 12 of them, which by the Central* Limit Theorem is approximately Gaussian, subtract the mean value (6), and obtain approximately Gaussian numbers with unit variance.

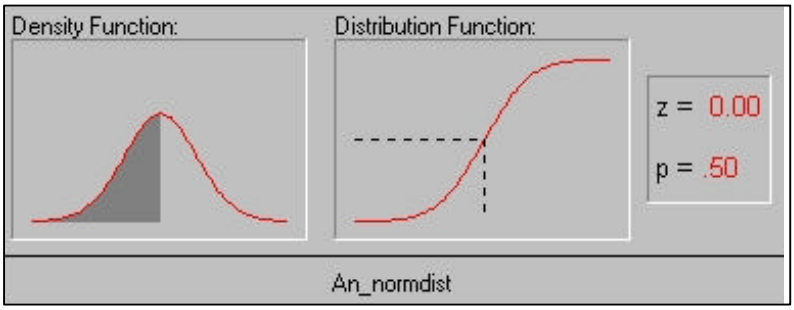


Figure 6. Normal probability density and distribution fcunctions [9]

Significance tests are based on certain assumptions: The data have to be random samples out of a well defined basic population and one has to assume that some variables follow a certain distribution, in most cases the normal distribution is assumed.

Power of a test is the probability of correctly rejecting a false *null hypothesis*. A null hypothesis is a hypothesis of no difference. *If a null-hypothesis is rejected when it is actually true, a Type I error has occurred***.** This probability is one minus the probability *of making a Type II error (b), which is the error that occurs when an erroneous hypothesis is accepted***.** Decreasing the probability of making a Type I error will increase the probability of making a Type II error. The probability of correctly retaining a true null hypothesis has the same relationship to Type I errors as the probability of correctly rejecting an untrue null hypothesis does to Type II error.

Anytime one test whether a sample differs from a population or whether two sample come from 2 separate populations, there is the assumption that each of the populations has its own mean and standard deviation. The distance between the two population means will affect the power of our test.

It should notice that what really made the difference in the size of b is how much overlap there is in the two distributions. When the means are close together the two distributions overlap a great deal compared to when the means are farther apart. Thus, anything that effects the extent the two distributions share common values will increase b (the likelihood of making a Type II error).

Many statistical methods are based on the assumption that data are normally distributed. If an initial histogram plot indicates that the data to be analysed may be normally distributed, we can perform another quick test, before conducting more formal tests (e.g. a chi-square test). In order to determine if a sample of data may have come from a Normal population, the best-fit Normal distribution is computed & compared with a histogram (Figure 7). The non-standardised variable *x* is plotted and the area under the curve is equal to the total frequency times the histogram class interval.

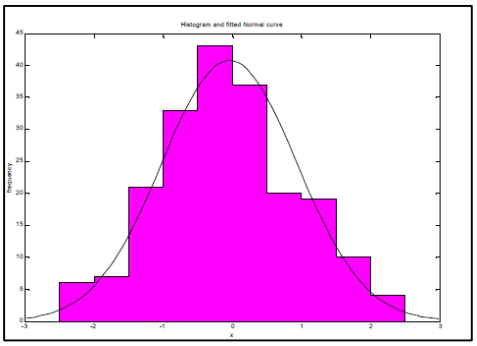


Figure 7. A histogram of 200 values drawn from a normal population with a mean 0 and standard deviation 1, together with the fitted Normal curve [9]

Correlation is a measure of the relation between two or more variables [10]. The measurement scales used should be at least interval scales, but other correlation coefficients are available to handle other types of data. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negatice correlation while a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents a lack of correlation. The most widely-used type of correlation coefficient is Pearson *r*, also called linear or product-moment correlation.

Pearson correlation (hereafter called *correlation*), assumes that the two variables are measured on at least interval scales, and it determines the extent to which values of the two variables are "proportional" to each other. The value of correlation (i.e., correlation coefficient) does not depend on the specific measurement units used. *Proportional* means *linearly related*; that is, the correlation is high if it can be "summarized" by a straight line (sloped upwards or downwards).

This line is called the *regression line* or *least squares line*, because it is determined such that the sum of the *squared* distances of all the data points from the line is the lowest possible. Note that the concept of *squared* distances will have important functional consequences on how the value of the correlation coefficient reacts to various specific arrangements of data. The correlation coefficient (*r*) represents the linear relationship between two variables. If the correlation coefficient is squared, then the resulting value (r2, the coefficient of determination) will represent the proportion of common variation in the two variables, i.e the "strength" or "magnitude" of the relationship). It is important to know this magnitude or "strength" as well as the significance of the correlation.

3. Research Results

Some mathematical examples will be performed using MATLAB so show its many varied functions and use.

Normal Disrtibution

Grades of chip samples from a body of ore have a normal distibution with a mean of 12% and a standard deviation of 1.6%. Find the probability of the grade of a chip sample taken at random will have a grade of:

1) 15% or less
2) 14% or more
3) 8% or less
4) between 8% and 15%

```
% mean 12%, sd 1.6%
m1  = 12;
sd1 = 1.6;
%prob <15%
%first standardise
z1 = sdiz(15,m1,sd1);
%then use (just created) cumulative prob fn
prob1 = cump(z1);

%  Now prob>14  %
z2     = sdiz(14,m1,sd1);
prob2  = 1-cump(z2);

%  prob<8       %, not problem here if z is negative
z3    = sdiz(8,m1,sd1);
prob3 = cump(z3);

% 8 % < prob<15 %
% already standerised 8 & 15  % (ie z3 and z1)
prob4 = cump(z1)-cump(z3)
```

Basic sample statistics

The following data are diameters (in mm) of clasts from a conglomerate:

23 24 27 29 29 30 33 33 34 38 45 60 60 88 126 221 256

```
load ex_2_2.dat   %loads file into array called ex_2_2

median1 = median(ex_2_2)
```

```
mean1   = mean(ex_2_2)

%Now to get geometric mean
l=1.0;
for I = 1:length(ex_2_2)
    x = ex_2_2(i);
    t = x*l;
    l = t
end
geom = l^(1/length(ex_2_2))
```

Poisson distribution

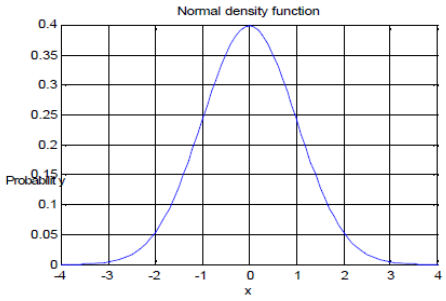
The number of major floods occurring in 50-year periods in a certain region has a Poisson distribution with a mean of 2.2. What is the probability of the region experiencing

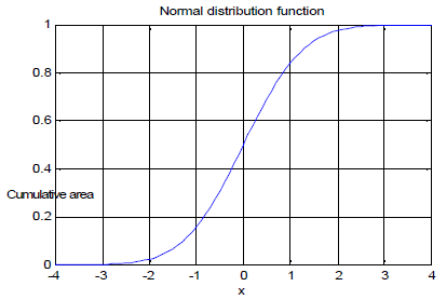
1) exactly one flood in a 50-year period?
2) Exactly one flood in a 25-year period?
3) At least one flood in a 50-year period?
4) Not more than two floods in a 25-year period?

```
% For a Poisson distribution
% Pr(X=x) = exp(-mean*t)*((mean*t)^x)/x!
% Nb 1 time period (t=1) is 50 years

% 2 floods in 50 years, X=2 t=1
m1 = 2.2

prob1 = exp(-m1)*((m1).^2)/factorial(2)
% 1 flood in 25 years, X = 1, t = 0.5

prob2 = exp(-m1*0.5)*((m1*0.5).^1)/factorial(1)
% at least one flood in 50 years (t=1) is all but zero in 50 years
% ie 1-P(X=0)

prob3 = 1-exp(-m1)*((m1).^0)/factorial(0) %nb 0!=1
% Not more than 2 is P(0)+P(1)+P(2), t=0.5

prob4 = (exp(-m1*0.5)*((m1*0.5).^0)/factorial(0))+(exp(-m1*0.5)*((m1*0.5).^1)…
/factorial(1))+(exp(-m1*0.5)*((m1*0.5).^2)/factorial(2))
```

Normal density function and cumulative density function

```
clear
x = [-4:0.1:4];
dens = dnorm(x);
dist = pnorm(x);
subplot(1,2,1)
plot(x,dens)
grid on
title('Normal distribution function')
xlabel('x')
ylabel('Probability')
subplot(1,2,2)
plot(x,dist)
grid on
title('Normal density function')
xlabel('x')
ylabel('Cumulative area')
```





Confidence intervals of means

1) Given data on the percentage of quartz in thin sections from an igneous rock, what is the confidence interval around the estimated mean quartz percentage in the rock?

```
qz =[23.5 16.6 25.4 19.1 19.3 22.4 20.9 24.9];
m1 = mean(qz);
v1 = var(qz);
n  = length(qz);
% standard error: root(s^2/n)
s1 = sqrt(v1/n);
% Now t(0.025,n-1) is prob not exceeded by 0.975
t1=qt(0.975,n-1);
% Confidence limits
c1 = m1-t1*s1
c2 = m1+t1*s1
```

2) Is there any evidence that two brachiopod samples could have been derived from populations having the same mean (data are contained in data files ex_2_22_a.dat and ex_2_22_b.dat)?

```
ex_2_22_a.dat;
load ex_2_22_b.dat;

% Calculate mean, std, variance, and sample size
m1 = mean(ex_2_22_a)
m2 = mean(ex_2_22_b)
s1 = std(ex_2_22_a)
s2 = std(ex_2_22_b)
ss1 = s1*s1
ss2 = s2*s2
n1 = length(ex_2_22_a)
n2 = length(ex_2_22_b)

% Calculate combined variance
sc=((n1-1)*ss1+(n2-1)*ss2)/(n1+n2-2)

%t dist with n1+n2-2=16 df; will be exceeded by 0025::0.975 will not be %
exceeded
t1 = qt(0.975,16)
st = t1*sqrt(sc*((1/n1)+(1/n2)))

%now difference in means...
dm=m1-m2
%Confidence limits
c1 = dm-st
c2 = dm+st
```

T-test

A random sample of 12 observations is obtained from a normal distribution. What value of the *t*-statistic will be exceeded with a probability of 0.025? The number of degrees of freedom (*df*) is n-1=11. Use the help function to find out how *dt*, *pt* and *qt* work.

```
% Using t-distribution generator (stixbox functions dt, pt and qt).
```

```
% t-stat will be exceeded by P=0.025 and is not exceeded by P=0.975

p  = 0.975;
df = 11; %n-1

%inverse t is qt
t = qt(p,df)
```

A random sample of 8 hand specimens of rock was analysed for organis material; th4 sample mean was found to be be 5.8 % and the sample standard deviation was 2.3. Does someone think it reasonable to suppose that the organix content of the rock is 5.0%?

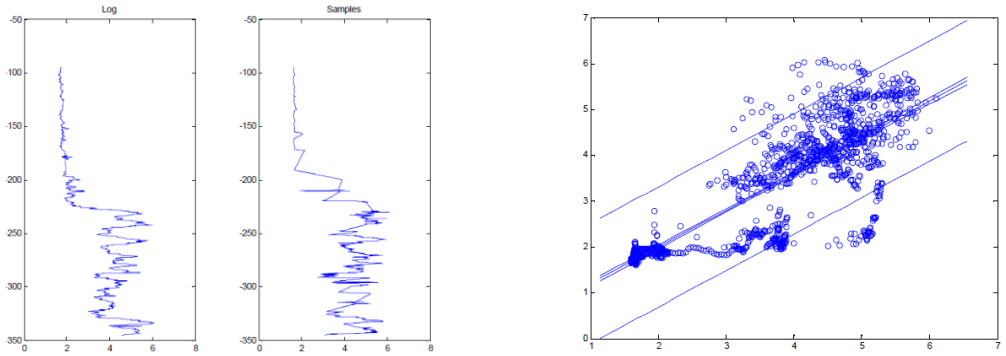
```
%Use function tstat

function t = tstat(m1,m2,s,n)
%Calculates t- test statistic
t = (m1-m2)/(s/sqrt(n))

m1 = 5.8; %sample mean
m2 = 5.0; %suggested mean
s = 2.3; %std
n = 8;
t2 = tstat(m1,m2,s,n)

%Now for 7df want t0.05 (ie not exceeded by 0.95)
t3 = qt(0.95,7)
%ie no reason to doubt mean (ie t2 < t3)
```

Linear and polynomial regression

On ODP Leg 183 (Kerguelen Plateau) sediment velocity data were collected based on both downhole geophysical logs and on laboratory measurements of core samples. The question arises: How well do the two sets of measurements correlate? One can only work with values collected below 80 m below sea floor, because the hole was cased above this depth, preventing us to collect data based on downhole logs.

```
% Velocities from Kerguelen ODP Leg183
clear
load vel_log.dat
load vel_samp.dat

deplog = vel_log(:,1);
depsamp = vel_samp(:,1);

vellog = el_log(:,2);
velsamp = vel_samp(:,2);

plot(vellog,-deplog,'o')
hold on
plot(velsamp,-depsamp,'+r')
title('Velocities from logs (blue) and from samples (red)');
xlabel('Velocity [km/s]')
ylabel('Depth [m]')
hold off

% Usable depth interval
depi = (95:0.15:345);
% Resample both sets of measurements at log sampling rate
% logs are more coarsely sampled than core data
help interp1
vell = interp1(deplog,vellog,depi);
vels = interp1(depsamp,velsamp,depi);
% Plot resampled data (velocity versus depth)
% on scatterplot
```

```
figure
plot(vell,vels,'+');

% Find out options for linreg stixbox function
help linreg

linreg(vell,vels)
```



4. Conclusion

This study has shown the different way in which of Matlab software in the teaching of the various topics such as normal distribution, confidential interval of mean, and linier regression. Matlab software usage is proposed to increase the understanding of these difficult topics among the undergraduate school students. It is therefore possible that with good course design, lecturers can have some degree of control over what topics, that Matlab software can be effective for improving performance in Statistics.

Acknowledgment

The author would like to thank to Head of LPPM ITNY, Dr. Hj. Ani Tjitra Handayani, for providing necessity services and Dr. Sugiarto for helping to furnish the paper.

References

[1] Snee, R. 1993. What's Missing in Statistical Education? *The American Statistician, 47(2)*, 149-154.
[2] Garfield, J. and Ben-Zvi, D. 2007. How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics, *International Statistical Review 75(3)*, 372-396, doi:10.1111.
[3] Watson, J. M. 1997. Assessing Statistical Thinking Using the Media, In *The Assessment Challenge in Statistics Education*, Gal, I. and Garfield, J. B. (Eds.). Amsterdam: IOS Press and The International Statistical Institute, 107-121.
[4] Nolan, D. and Speed, T. P. 1999. Teaching Statistics Theory through Applications. *The American Statistician*, 53, 370-375.
[5] Ogunkunle, R. A. and Charles-Ogan, G. 2013. Dependence on calculators for Acquisition of Basic Skills in Junior Secondary School Mathematics. *Journal of Research in National Development (JORIND).11(1), 228-232.*
[6] Abdullah, K. A. Hashim, N. and Yusof, Z. 2010. *The Development of Computer-aided Learning for Computer Numerical Control Machine: A plot study*, In: 2nd International Congress for Engineering Education(ICEED), 94-99, ISBN: 978-1-4244-7308-3
[7] Attaway, S. 2012 . MATLAB. *A Practical Introduction To Programming and Problem Solving. 2nd ed.* Boston; Elsevier Inc. USA.
[8] Matlab® Ver. R2017a. Available: http://www.mathworks.com
[9] Baglivo, J. A. 2005. *Mathematica Laboratories for Mathetical Statistics. Emphasizing Simulation and Computer Intensive Methods*. ASA-Siam Series on Statistics and Applied Probability. SIAM Philadelphia, ASA, Alexandria, VA.
[10] Sanbhag, D. N., and Rao, C. R. 2003. *Hanbook of Statistics*. Vo. 21. Elsevier. Amsterdam.