Mathematical Statistics

SF1901 Probability Theory and Statistics: Autumn 2016
Lab 2 for
TCOMK

Introduction

This is the instructions for Computer exercise 2, please bring a printed copy to computer exercise 2. Please read the instructions carefully, and make sure that you understand what the MATLAB code included does. The computer exercise is pass/fail. In order to pass you first have to be able to present written solutions to the preparatory exercises. Each student should have prepared solutions **individually**. For the computer exercise it is allowed (and encouraged) to work in groups of **at most two** persons per group. If you pass the computer exercise you will be given 3 bonus points at the written exam Wednesday October 26 2016, **given that you at the exam obtain at least 20 points without the bonus points**.

1 Preparatory exercises

1. Plot the density functions for the following distributions
   (a) $N(0,1)$, $N(-1,10)$, $N(100,0.01)$.
   (b) Exp(1), Exp(2), Exp(10).
   (c) $\Gamma(1,2), \Gamma(5,1)$.
2. State some properties which are characteristic for data coming from a normal distribution.

Comments: .....
.....
.....

3. State a $1-\alpha$ confidence interval for μ when the data consists of observations of n independent $N(\mu, \sigma)$ distributed random variables, when σ is known/unknown.

   Comments: .....
   .....
   .....

4. Define the likelihood function, the log-likelihood function, and explain the relationship between the two. Describe the idea behind the method of least squares (LS) and the method of maximum likelihood (ML), respectively.

   Comments: .....
   .....
   .....
   .....

5. When a random variable X has the density function

$$f_X(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}, \qquad x \geq 0.$$

   then it is said to be Rayleigh distributed. Assume that you have observed outcomes of n independent Rayleigh distributed random variables.

   - Determine the ML estimate of the parameter b.

     Answer: .....
   - Determine the LS estimate of the parameter b.

     Answer: .....

6. Derive a confidence interval for the parameter b with approximate confidence level $1-\alpha$. Motivate when and why it is reasonable to make the approximation.
   Hint: Base the interval on the LS estimate.

   Comments: .....
   .....
   .....
   .....

7. Describe the idea behind linear regression. Explain what polynomial regression is. The following link may be of use:

   https://en.wikipedia.org/wiki/Polynomial_regression

   Comments: .....
   .....
   .....
   .....

8. Describe how the MATLAB command `regress` can be used to obtain estimates of the parameters in the following model

$$w_k = \log(y_k) = \beta_0 + \beta_1 x_k + \varepsilon_k \tag{1}$$

9. Explain the idea behind bootstrap. The following link may be of use

   https://en.wikipedia.org/wiki/Bootstrapping_(statistics).

Further introduction

Start by downloading the following files

- `wave_data.mat`
- `resistorer.mat`
- `moore.mat`
- `poly.mat`
- `birth.dat`
- `birth.txt` - description of the data `birth.dat`

from the homepage of the course. Make sure the files are downloaded to the directory you will be working in. To make sure the files are in the right directory you can type `ls *.*at` to list the files.
You can write your commands directly at the prompt in MATLAB, but usually it is easier to work in the editor. If the editor is not open you can open it and create a new file by typing `edit lab2.m`.

Problem 1 - Maximum likelihood/Least squares

The code below generates observations from a Rayleigh distribution and plots the estimate `my_est`. Use the two point estimates that you derived in the preparatory exercise 5 (to plot `my_est_ls`, type the formula for it and and set `my_est` equal to `my_est_ls`).

```
%% Problem 1: Maximum likelihood/Least squares
M = 1e4;
b = 4;
x = raylrnd(b, M, 1);
hist_density(x, 40)
hold on
my_est_ml = % Write the formula for your ML estimate here
my_est_ls = % Write the formula for your LS estimate here
my_est=my_est_ml;
plot(my_est, 0, 'r*')
plot(b, 0, 'ro')
hold off
```

Do your estimates look good?

Comments: .....
.....
Check what the density function looks like by plotting it along with your estimate:

```
plot(0:0.1:6, raylpdf(0:0.1:6, my_est), 'r')
hold off
```

Problem 2 - Confidence intervals

In this section you should investigate data from a Rayleigh distributed signal; you should estimate the parameter for the distribution and determine a confidence interval for the parameter. Load the data by typing `load wave_data.mat`. The file contains a signal that you can plot by typing the following

```
%% Problem 2: Confidence interval
load wave_data.mat
subplot(211), plot(y(1:100))
subplot(212), hist_density(y)
```

If you change `y(1:100)` to `y(1:end)`, then you can see the whole signal. Estimate the parameter based on the observations found in `wave_data.mat` in the same way that you did in the previous problem. Assign your estimate to `my_est`. Compute a confidence interval for the parameter and assign the upper and lower limits to `upper_bound` and `lower_bound`, respectively (recall the preparatory exercise 6). Write down the results:

Answer: .....
.....

. . . . . .
Now plot the confidence interval for the parameter

```
% ...
hold on     % holds the current plot
plot(lower_bound, 0, 'g*')
plot(upper_bound, 0, 'g*')
```

Check what the density function looks like by plotting it along with your estimate, just as you did in the previous problem.

```
% ...
plot(0:0.1:6, raylpdf(0:0.1:6, my_est), 'r')
hold off
```

Does it look as if though the distribution fits the data?

Answer: . . . . . .

The Rayleigh distribution can be used to describe the fading of a radio signal. Experimental work in Manhattan has shown that the effect of the densely built city on the propagation of a radio signal can be approximated by Rayleigh fading. [1].

Problem 3 : Simulation of confidence intervals

A $1-\alpha$ confidence interval for the parameter μ covers the true (unknown) μ with probability $1-\alpha$. The purpose of this problem is to give an understanding of the concept of confidence level by looking at simulations. The code on yhe next page uses $n = 25$ independent observations from the $N(2, 1)$ distribution to compute a confidence interval for the expectation with confidence level 95% (we pretend to forget that we know what the true value is). If this is repeated 100 times leaving us with 100 confidence intervals, how many do we expect to cover the true parameter?

What do the horizontal lines and the vertical line indicate? How many intervals cover the true value of μ? Does the result agree with what you expected? Run the simulations a few more times and interpret the results.

Answer: . . . . . .
. . . . . .
. . . . . .

```
%% Simulation of confidence intervals
% Parameters:
n = 25; % Number of measurements
mu = 2; % Expected value
sigma = 1; % Standard deviation
alpha = 0.05;

%Simulation of n * 100 observations. (n observations for ...
    each interval and 100 intervals)
x = normrnd(mu, sigma,n,100); %n x 100 matrix of observations

%Estimation of mu by mean
xbar = mean(x); % vector containing 100 means.

%Computation of upper and lower limits
lowerl = xbar - norminv(1-alpha/2)*sigma/sqrt(n);
upperl = xbar + norminv(1-alpha/2)*sigma/sqrt(n);

%Plot all the intervals making the ones which do not cover ...
    the true value red
figure(1)
hold on
for k=1:100
    if upperl(k) < mu
        plot([lowerl(k) upperl(k)],[k k],'r')
    elseif lowerl(k) > mu
        plot([lowerl(k) upperl(k)],[k k],'r')
    else
        plot([lowerl(k) upperl(k)],[k k],'b')
    end
end
%b1 and b2 are only used to make the figure look nice.
b1 = min(xbar - norminv(1 - alpha/2)*sigma/sqrt(n));
b2 = max(xbar + norminv(1 - alpha/2)*sigma/sqrt(n));
axis([b1 b2 0 101]) % Minimizes amount of unused space in ...
    the figure

%Plot the true value
plot([mu mu],[0 101],'g')
hold off
```

Now change the values for μ, σ, n and α (one at a time). How does changing the parameters affect the results?

Answer: .....
.....
.....

Problem 4 : Fitting a distribution

Load `resistorer.mat` and study the data (which describes a measured property in a number of resistors) using a histogram. Also examine the data using the command `normplot`. Which distribution do you think the observations come from? Can you rule out any distributions? Can you think of a reason to be interested in the distribution of a particular property of resistors?

Comments: .....
.....

Problem 5a: Linear regression

In this problem you should look at the phenomenon known as Moore's law. Load the data from `moore.mat` in the same way as before. In the data `y` represents the number of transistors/unit area whereas `x` is the year. This means that if you plot `y` against `x` then what you see is a plot of the development over time of the number of transistors/unit area.

Let us introduce the following model

$$w_i = \log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{2}$$

Now estimate the parameters β_0 and β_1 using MATLAB's function `regress`. If you use the data from 1971 until 2011 to estimate the parameters, what is your prediction for the number of transistors the year 2020?

Answer: .....

Problem 5b: Polynomial regression

Start by loading the file `poly.mat`. Then plot `y1, y2, y3`, separately against `x1, x2` and `x3`, respectively. Do the plots look as if the could be described by a polynomial?

Comments: .....
.....

Let us introduce the following model

$$y_k = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n. \tag{3}$$

Now, for each of the three data sets, you should form an appropriate matrix X. The matrix X should represent a polynomial of a degree that seems appropriate for the data set. For the model (3) above the matrix X would

be formed in the following way

$$X = \begin{bmatrix} 1 & x & x^2 & \dots & x^n \\ 1 & x & x^2 & \dots & x^n \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x & x^2 & \dots & x^n \end{bmatrix}. \tag{4}$$

Having formed your matrix X you should use it to find an estimate $\hat{\beta}$ of your parameters using `regress`, and then plot your estimated model

$$\hat{y} = X\hat{\beta}. \tag{5}$$

In the case of `y1`, you would have:

```
%% Problem 4: Regression
y_hat = X*beta_hat;
plot(y1, '.')
hold on
plot(y_hat, 'r.')
hold off
```

Now plot the residuals

```
res = y_hat - y1;
subplot(211), normplot(res)
subplot(212), hist(res)
```

Which distribution do the residuals seem o be taken from?

Answer: .....

Does it seems reasonable to use polynomial regression on the data found in the file `poly.mat`?

Answer: .....

Linear regression was developed in the late 18th century by the young Gauss. The method gained attention when it was used to successfully predict the orbit of the first discovered asteroid, Ceres. Linear regression is today in extensive use with applications in virtually all sciences dealing with data.

The theory is treated in more detail in the course "Regression Analysis".

Problem 6 - Bootstrap of the difference in mean to estimate the difference in expectations

You should now study the difference in expected values in two populations. For instance the difference in expected birth weight between children whose mothers smoked during the pregnancy, and children whose mothers did not smoke during pregnancy (if you like you can take two other populations and/or some other variable to study).
In the file `birth.txt` you can see that column 20 of `birth.dat` contains information about smoking habits during the pregnancy. The values 1 and 2 indicate that the mother did not smoke during the pregnancy, whereas the value 3 indicates that the mother did smoke during the pregnancy. You can therefore create two vectors `x` and `y` containing birth weights for children of non-smoking and smoking mothers, respectively, using the following code

```
>> x = birth(birth(:, 20) < 3, 3);
>> y = birth(birth(:, 20) == 3, 3);
```

The code `birth(:, 20) < 3` returns a vector of "true" (indicated by the value 1) and "false" (indicated by the value 0), and only those rows of column 3 (which contains the birth weights in `birth`) for which the comparison is true, end up in the vector `x`. Use the function `length` or the command `whos` to find out the sizes of the vectors `x` and `y`.
To estimate the difference between the expected values in the two populations you can use the difference between the mean values.

```
mean(x) - mean(y).
```

To get an idea of the uncertainty in this estimate you should use bootstrap and simulate M bootstrap samples and compute the difference between the means on the new samples according to

```
>> thetaboot = bootstrp(M, @mean, x) - bootstrp(M, @mean, y);
```

Does it look as if the differences between the means are taken from a normal distribution? (Do a histogram over `thetaboot` to get an idea of the distribution of the sample variable.) What does your confidence interval for the difference in expectations θ look like if you use the following

```
>> quantile(thetaboot, [0.025, 0.975])
```

What does the interval look like if you instead use the method found in the textbook for the difference between expected values (means)?

Answer: .....

Comments: .....
.....

Referenser

[1] Chizhik, Dmitry and Ling, Jonathan and Wolniansky, Peter W and Valenzuela, Reinaldo A and Costa, Nelson and Huber, Kris (2003). Multiple-input-multiple-output measurements and modeling in Manhattan *Selected Areas in Communications, IEEE Journal on*, Vol **21**, p. 321-331.