MATH2099/MATH2859
Probability, Statistics and Information

Statistics Laboratory Class Week 1 : Introductory Matlab Tutorial

In your statistics course, there are scheduled computer tutorials in which you explore the use of Matlab in statistics. Matlab is available on both the Linux and the Windows desktops in the labs in the School of Mathematics and Statistics in the Red Centre (on the ground floor and on the mezzanine level). Each student has a scheduled class with a tutor in the labs approximately every second week for the session. Students are welcome to use the labs at any time when they are not busy for other classes. Check notice boards for the current opening hours. **You do need to be able to use Matlab, so start as early as you can in the session!**

Exercises and data sets for the computer tutorials will be made available on your course web page (Moodle). Solutions to these exercises will usually be available from the course web page the following week. There is a general 'Introduction to Matlab' document available through the School of Mathematics and Statistics web page also available from your course web page. In addition there is an 'Introduction to Matlab and Statistics Using Matlab' (referred to here as **SUM**) document which is available from the course web page. **In all Matlab tutorials make sure you understand the output you obtain.** When you produce a graph, look at it! What does it tell you about the data? Ask you tutor if you have questions. Explore the possibilities of Matlab - you may find it very useful in all your courses!

Instructions

Pay particular attention to section 2 on plotting and graphing, and to section 3.1 where you use the Import Wizard to ***input the data files*** *for the computer tutorials in this course.*

Double click on the Matlab icon and a Matlab Command Window should pop-up. The `>>` is a prompt, requiring you to enter commands. One of the first commands you should type in is to find out what your working directory is. The working directory is where you will save the results of your calculations. So, enter

```
>> pwd
```

Note that `pwd` stands for 'print working directory'. You probably would get an output like :

```
ans = /student/your last digit of your student id number/username
```

Matlab always stores the result of its last calculation in a variable called `ans` (short for answer). Now this indicates the working directory.

Online help can be accessed for all Matlab commands by issuing the `help` command, or you can use `helpwin`, `helpdesk`, or `demo`.

>> help ⟨*type name of command here*⟩

To get started, you can simply type `>> help`

1. Variables, Vectors and Matrices

Variables in Matlab are like variables in any other programming language (C, C++ etc.), however, one difference is that you do not have to define them by indicating the type. Also, variable names are case sensitive and can be used to refer to a single number (a **scalar**) or an array of numbers (a **matrix** or a **vector**). To create a row vector in Matlab, for example (note the spaces):

```
>>  r = [1 2 3 4]
r =
 1 2 3 4
```

A column vector can be created, for example, (note semicolons):

```
>> c = [1; 2; 3; 4]
c =
 1
 2
 3
 4
```

On the other hand, you can use the ' operator (transpose)

```
>> c = r'
c =
 1
 2
 3
 4
```

Vectors can be created by incrementing a starting value with a constant quantity. For example (note the use of colons):

```
>> r = [0:2:10]
r =
 0 2 4 6 8 10
```

This creates a row vector, with the first element = 0 and each element incremented by 2; until the final value of 10. You can index specific parts of a vector. For example, to get the third element in the vector r:

```
>> r(3)
ans =
 4
```

Matrices are 2 dimensional quantities and are created similarly. Eg:

```
>> a = [1 2 3; 4 5 6; 7 8 9; 10 11 12]
a =
 1  2  3
 4  5  6
 7  8  9
 10 11 12
```

Matrix a is a 4 × 3 matrix (4 rows and 3 columns). We can also use the incrementation principle, for example:

```
>> b = [0:2:10; 1:2:11]
b =
 0 2 4 6 8 10
 1 3 5 7 9 11
```

Matrix b is a 2 × 6 matrix. Entries of the matrix, for instance the entry in the 2nd row, 5th column can be accessed using the notation:

```
>> b(2, 5)
ans =
 9
```

To extract data, for example the first column or second row of the matrix, use colons:

```
 >> a(:,1)
ans =
     1
     4
     7
    10
```

(so this gives us the first column in the matrix a)

```
  >> a(2,:)
ans =
     4     5     6
```

(so this gives us the second row in the matrix a)

1.1. **Vector and Matrix Operations.** The basic arithmetic operations `+, -, *, /` can be used for vectors and matrices. These would generate corresponding output vectors or matrices. For example, to add two vectors:

```
>> a = [1 2 3 4];
>> b = [5 6 7 8];
>> c = a+b
c =
 6 8 10 12
```

NOTE THAT the semicolons (;) in the first two commands direct Matlab not to echo the values of the variables a and b on to the screen immediately after you type them. Check that you understand what this means! What is the difference when you repeat the previous three commands and don't put semicolons at the end of the commands? What if you do put a semicolon at the end of the command `c = a+b` ?

As you know from algebra, only vectors (or matrices) that have the same size can be added or subtracted. For example:

```
>> a = [1:3:20; 21:3:40];
>> b = [2:3:20; 22:3:40];
>> c = a-b
c =
 -1 -1 -1 -1 -1 -1 -1
 -1 -1 -1 -1 -1 -1 -1
```

Matrix multiplication using the * symbol is possible only if the number of columns in the first matrix equals the number of rows in the second:

```
>> a=[1 2 3; 4 5 6]
a =
 1 2 3
 4 5 6
>> b = a'
b =
 1 4
 2 5
 3 6
>> c=a*b
c =
 14 32
 32 77
```

If you want to multiply corresponding elements of two matrices (which have the same size), then you do an array multiply using the `.*` symbol. For example

```
>> a = [1 2 3 4; 5 6 7 8]; b=[2 2 2 2; 3 3 3 3];
>> c=a .* b
c =

     2     4     6     8
    15    18    21    24
```

2. **Plotting**

The `plot` command is used for generating 1-D (functions of one variable) plots. Do `>> help plot` for complete details. For example:

Let's make a graph of $y = \sin(x)$, for x on the interval $x = 0$ to $x = 10$.

```
>> x = [0:0.1:10]; (here .1 is the increment)
>> y = sin(x); (notice how the sin function operates on each element
 of the entire row vector x, to generate another row vector y)
>> plot (x, y)
```

Surface or 2-D (functions of two variables) plots are generated using the `surf` command. To clear a plot, type in `>> clf` (clear figure)

2.1. **New figure, and subplot commands.** To generate another plot window, do `>> figure`. If you want to have several plots on the same figure in a grid then you can use the `subplot` command. For example, to obtain a graph of $y = \sin(x)$, then on a second separate figure a graph of $y = \cos(x)$, then on a third separate figure, graphs of both functions side by side:

```
>> x = [0:0.1:10];
>> y1 = sin(x);
>> y2 = cos(x);
>> plot (x, y1)
>> figure
>> plot(x,y2)
>> figure
>> subplot(1,2,1)
>>  plot (x, y1)
>> title('graph of sin(x)')
>> subplot(1,2,2)
>> plot(x,y2)
>> title('graph of cos(x)')
```

In the `subplot` command the first two coordinates state the size of the grid (in this example, a 1×2 grid of graphs), and the third coordinate says where to put the next requested plot.

3. **File input/output**

In many instances, the simplest way to import data into MATLAB is to use the Import Wizard. All you need to do is click Import Data. In this course we only require you to be able to import data using the Import Wizard and this is the same whether you are running MATLAB on Linux or on Windows.

3.1. **Data files for these tutorials.** There are various data files which we will use in this course. These are available from the course web page (Moodle). You should now use the following instructions and use the Import Wizard (Home → Import Data) to check that you can import the following files without any problem.

Instructions: **For each data file**, locate the file on the course web page and then save it to your own directory on the computer. **Save the files as .txt files**. Exactly how you do this may depend on which browser you are using for the course web page. You may be able to simply save it to a file or you may find it easier to open a text file and cut and paste the data and then save it to your own directory. Experiment with the Import Wizard so that you understand how it works. Make sure you save the data files in a location you'll be able to easily access in the future, as we will make exhaustive use of them in these laboratories.

- **inflows.txt** This data file contains two columns of data, one called Year and one called Inflows. Select this data file and click **Import Data**. Alternatively, you can right-click on the data file and select **Import Data**. You should see that there are the two columns of data, with names at the top, and 25 pairs of values. The columns are separated by a tab so check that

Tab is selected at the top. You can edit the variable names as **Year** and **Inflow**. This will create vectors in MATLAB called `Year` and `Inflow`. Note that **the names are case sensitive**. Finally click the green tick **Import Selection**.

To list the data you can enter `>> Year`. (What happens if you forget the capital letter?) Then enter `>> Inflow`. To check you have all the data you might check the length of the vectors by `>> length(Year)`. (You should get 25.) To see the two columns you could enter `>> [Year Inflow]`. (What happens if you enter `>> [year;Inflow]`?)

- **kevlar90.txt** Locate this file and then save it to your own directory. Note there is only one column and there is no name at the top. Import it using the Import Wizard. How long is the column? The vector itself is called `kevlar90`. Check the length of the vector. (There should be 101 values).
- **shearstrength.txt**. Locate, save and import this data file. There is one column of data. How many values are there? After importing the values, check you can list them on the screen.
- **porevolume.txt** This data set contains just one variable in one column. Locate it, save it and import it using the Import Wizard.
- **fusiondat.txt** Locate, save and import this data file. There are two columns of data values.
- **abs.txt** The file abs.txt contains data on internal deadband impedances for front and rear sensors from an antilock braking system (ABS). Each row of the dataset corresponds to the same unit, with the first measurement being the front sensor impedance and the second the rear sensor impedance (in kohms). There are two columns. Import the dataset as it is: the whole data set matrix is called abs. (Try `>> abs`).
- **concrete.txt** Locate, save and import this data file. This data file has three columns of data.
- **rain.txt** The file rain.txt contains rainfall (first column) and runoff (second column) measurements at Pontelagoscuro on the Po river in northeast Italy, for the 31 years 1918 to 1948. Locate, save and import the data.
- **milk.txt** This data file has two columns, x = milk production (kg/day) and y = milk protein (kg/day) for a sample of Holstein Friesian cows. Locate, save and import the data.

3.2. **Saving data (and other ways of data input).** Alternatively, the **save** and **load** commands are used for saving data to disk or loading data from disk, respectively. To save values in a matrix or vector, called, for instance, `y`, do:

```
>> y = [1 2 3 4; 5 6 7 8];
>> save y.txt y -ascii
```

The -ascii option ensures that the data is saved in ASCII form, so that it can be read by other programs - Notepad, WordPad, Kwrite (in Linux) etc. Examine the file y.txt using one of these programs. The file is generated in the working directory. An ASCII data file dat.txt can be loaded using `>> load dat.txt` This provides a variable called `dat` in the MATLAB workspace. All manner of vector/matrix operations can be performed on `dat`, just like any other variable.

For learning all the `load/save` options use `>> help save` and `>> help load`.

4. **Executable files (m-files)**

Executable files in MATLAB are generated by storing a list of MATLAB commands in a file given the extension `.m`. These files are called **m-files** (or scripts). To create an m-file, use the New Script option. As an example, type in the following commands in the m-File Editor Window:

```
x = [0:0.1:10];
y = cos(x);
plot(x,y)
xlabel('x')
ylabel('y')
title('A plot of y=cos(x)')
```

Save the file using the Save Option in the m-File Editor Window. Call it, say, `cosineplot.m`. Now, to run this program in MATLAB, move over to the MATLAB Command Window and just type in `>> cosineplot` .

End of Introductory Laboratory Class (Week 1).

MATH2099/MATH2859
Probability, Statistics and Information

Statistics Laboratory Class Week 2 : Data and Histograms

Instructions

After this tutorial you should be able to: Access Matlab*, and use the help facilities; enter and manipulate vectors and matrices; transform data; enter data and display appropriately a histogram with appropriate titles and labels; comment appropriately on a graphical display of data; (draw a stemplot of data (on paper, not using* Matlab*); construct a histogram using the* `histogram` *command with the* Matlab *default bins or your own choice of bins.*

Browse the general 'Introduction to Matlab' document and the 'Statistics Using Matlab' (here referred to as ***SUM****) document to find extra information. Remind that you can also use the* `help` *command at any time to get information about any Matlab command.*

Do not just enter Matlab *commands! Make sure you understand the output you obtain from the commands you use. When you produce a graph, look at it! What does it tell you about the data? Ask you tutor if you have questions. Explore the possibilities of Matlab, go beyond what is actually required!*

Solutions to these exercises will be available from the course web page later.

Exercise 1*

The Hardap Dam is the largest dam in Namibia, with a surface area of approximately 25 square kilometres, a capacity of 320 million cubic metres, and a dam wall 862 metres long. Table 1 shows annual maximum floodpeak inflows to the dam, in cubic metres per second, for years 1962/63 to 1986/87. (Data from Metcalfe, Andrew V. (1997). Statistics in Civil Engineering. London: Arnold ; New York: Wiley).

Observation	*Year*	*Inflow*	*Observation*	*Year*	*Inflow*
1	1962	1864	14	1975	1506
2	1963	44	15	1976	1508
3	1964	146	16	1977	236
4	1965	364	17	1978	635
5	1966	911	18	1979	230
6	1967	83	19	1980	125
7	1968	477	20	1981	131
8	1969	457	21	1982	30
9	1970	782	22	1983	765
10	1971	6100	23	1984	408
11	1972	197	24	1985	347
12	1973	3259	25	1986	412
13	1974	554			

Table 1. Annual maximum floodpeak inflows to Hardap dam

a) Recover the vectors `Year` and `Inflow`, imported from the *inflows.txt* data set into Matlab in the first lab class.

b) Construct a frequency histogram of the inflows using the `histogram` command. Give the graph a title, and label the x- and y-axes appropriately. (Reference SUM p.7-9 or use the `help` command). Explore the `histogram` command so that you understand how to label appropriately.

c) Create a new data vector with values equal to the logarithms of the inflows. Construct a histogram of the new data vector with appropriate labels and titles.

d) The coefficient of skewness is a measure of asymmetry of the data. It is defined as

$$g = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$$

where s is the sample standard deviation. It gives an indication of departures from symmetry: a positive value of g may indicate a long “right tail” in the data with values above the median tending to be further away from the median than values below the median, while negative skewness may indicate a long “left tail” with values below the median tending to be more extreme.

Use MATLAB to calculate the skewness of the two datasets (see SUM p.6 or use help command). Compare the shapes of the histograms for the inflows and the log transformed inflows, and comment on the degree and type of skewness.

e) Without using MATLAB, construct a stem and leaf plot for the log-transformed inflows (round to the tenths place). Note that this is a paper-and -pencil exercise, but you can use MATLAB to help order the data using the `sort` command.

f) Compare the histograms and stem and leaf plots for the log transformed data. What information is there which is contained in one and not the other?

g) Write an **m.file** to carry out some or all of this exercise.

Exercise 2*

The table below shows the age at taking office for the 44 US presidents to date.

President	**Age**	**President**	**Age**	**President**	**Age**
Washington	57	*Lincoln*	52	*Hoover*	54
J.Adams	61	*A.Johnson*	56	*FDRoosevelt*	51
Jefferson	57	*Grant*	46	*Truman*	60
Madison	57	*Hayes*	54	*Eisenhower*	61
Monroe	58	*Garfield*	49	*Kennedy*	43
J.Q.Adams	57	*Arthur*	51	*L.B.Johnson*	55
Jackson	61	*Cleveland*	47	*Nixon*	56
VanBuren	54	*B.Harrison*	55	*Ford*	61
W.H.Harrison	68	*Cleveland*	55	*Carter*	52
Tyler	51	*McKinley*	54	*Reagan*	69
Polk	49	*T.Roosevelt*	42	*G.W.H.Bush*	64
Taylor	64	*Taft*	51	*Clinton*	46
Fillmore	50	*Wilson*	56	*G.W.Bush*	54
Pierce	48	*Harding*	55	*Obama*	47
Buchanan	65	*Coolidge*	51		

Table 2. Age at taking office for the 44 US presidents to date

a) Recover the vectors `name`, `age` and `country`, imported from the *presdat2012.txt* data set into MATLAB in the first lab class. Extract the ages of the US presidents only. *Hint:* the command `strcmp` is used to compare character strings. The command `A=age(strcmp(country,'US'))` will create a vector `A` that contains `age` values for which `country='US'`.

b) Construct a titled and labelled plot of president age versus time order. Comment on any temporal patterns you see (i.e., what is the pattern, if any, over time).

c) Construct a titled and labelled frequency histogram of president ages.

d) Use the `histogram` command to construct an appropriately labelled frequency histogram showing numbers of presidents whose ages fall in the ranges 42-44, 45-47, 48-50, ...,69-71. *Hint:* enter `help histogram` for information on this command. You need to set up an edges vector to use in the `histogram` function.

e) Compare the two frequency histograms produced in the previous steps. Comment briefly on the visual impact of the choice of class intervals used in both.

f) Write an **m.file** to carry out some or all of this exercise.

End of Laboratory class for Week 2.
If you are unable to finish during the class then try to finish in your own time before your next lab class so that you can seek help then if needed.

MATH2099/MATH2859
Probability, Statistics and Information

Statistics Tutorial Class Week 3 : Descriptive Statistics

Instructions

The following are exercises that will covered during your tutorial class.

Note that ***not every topic in the course can be covered by the tutorial exercises.*** *You are expected to take an active part in the learning process by working by yourself on most of the topics. In particular,* ***you are expected to attempt these tutorial questions beforehand****. During the tutorial, the tutor will go through the answers to some of the questions while directing explanation to areas where students indicate they have difficulty.*

Solutions to these exercises will be available from the course web page at a later date.

Exercise 1*

a) Answer the following yes/no questions, and **explain** your answer.
   (i) Will the sample mean always correspond to one of the observations of the sample?
   (ii) Will exactly half of the observations in a sample always fall below the mean?
   (iii) Will the sample mean always be the most frequently occurring data value in the sample?
   (iv) Can the sample standard deviation be equal to zero?
   (v) Can the sample median be equal to the sample mean?

b) (i) Suppose that you add 10 to all of the observations in a sample. How does this change the sample mean? How does it change the sample standard deviation?
   (ii) Suppose that you multiply all of the observations in a sample by 2. How does this change the sample mean? How does it change the sample standard deviation?
   (iii) A sample of temperature measurements in a furnace yielded a sample average of 446°Celsius and a sample standard deviation of 5.8°Celsius. You would like to communicate this information to an American colleague. What are the sample average and the sample standard deviation expressed in °Fahrenheit? (*Hint :* temperature in °C = (temperature in °Fahrenheit - 32)×5/9)

Exercise 2*

An experiment to investigate the survival time (in hours) of an electronic component consists of placing the parts in a test cell and running them for 100 hours under elevated temperature conditions (this is called an 'accelerated life test'). Eight components were tested with the following resulting failure times :

$$75 \ 63 \ 100^+ \ 36 \ 51 \ 45 \ 80 \ 90$$

The observation 100^+ indicates that the unit still functioned at 100 hours. Is there any meaningful measure of location that can be calculated for these data? What is its numerical value?

Exercise 3

Consider a sample of observations $x_1, x_2, \ldots, x_n$. For what value a is the quantity $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - a)^2$ minimised? Interpret in terms of the location and dispersion parameters that you know.

Exercise 4*

The following data is a sample of shear strength, (MPa) of a joint bonded in a particular manner :

$$22.4, 40.4, 16.4, 73.7, 36.6, 109.9, 30.0, 4.4, 33.1, 66.7, 81.5$$

a) Determine the 5-number summary.

b) Determine the iqr. Are there any outliers (by the 1.5× iqr rule)?

c) Construct a box-plot and comment on its features.
d) Determine the mean $\bar{x}$ and the sample standard deviation s.
e) By how much could the largest observation be decreased without affecting the iqr?

Exercise 5

a) The female students in an undergraduate engineering course self-reported their heights (in inches) to the nearest inch. The data are :

62 64 66 67 65 68 61 65 67 65 64 63 67 68 64 66 68 69 65 67 62 66 68 67 66 65 69 65 70 65 67 68 65 63 64 67 67

Calculate the sample mean, the sample standard deviation and the sample median of height.

b) In the same class, the male students self-reported their heights as follows :

69 67 69 70 65 68 69 70 71 69 66 67 69 75 68 67 68 69 70 71 72 68 69 69 70 71 68 72 69 69 68 69 73 70 73 68 69 71 67 68 65 68 68 69 70 74 71 69 70 69

Calculate the sample mean, the sample standard deviation and the sample median of height.

c) Construct a comparative stem-and-leaf plot for both genders. Comment on any important features that you notice in this display.

Exercise 6

Direct evidence of Newton's universal law of gravitation was provided from a renowned experiment by Henry Cavendish (1731-1810). In the experiment, masses of objects were determined by weighting, and measured force of attraction was used to calculate the density of earth. The values of the earth's density estimated by Cavendish, expressed as a multiple of the density of water (1 g/cm^3), are:

5.50 5.30 5.47 5.10 5.29 5.65 5.55 5.61 5.75 5.63 5.27 5.44 5.57 5.36 4.88 5.86 5.34 5.39 5.34 5.53 5.29 4.07 5.85 5.46 5.42 5.79 5.62 5.58 5.26

(*Source :* Philosophical Transactions, 17 (1798), 469)

a) Find the sample mean, the sample standard deviation and the sample median of these data.
b) Determine the iqr. Are there any outliers (by the 1.5× iqr rule)?
c) Construct a box-plot and comment on its features.
d) Would you suggest the sample mean or the sample median as single estimate of the density of earth from Cavendish's data?

Exercise 7

A.A. Michelson (1852-1931) made many series of measurements of the speed of light. Using a revolving mirror techniques, he obtained

12 30 30 27 30 39 18 27 48 24 18

for the differences (velocity of light in air) - 299,700 km/s. (*Source :* The Astrophysical Journal, 65 (1927), 11.)

a) Draw a dotplot.
b) Find the median and the mean. Locate both on the dotplot.
c) Find the variance and standard deviation.
d) Find the quartiles.
e) Find the minimum, maximum, range, and interquartile range.

Exercise 8

An experimental study of the atomisation characteristics of biodiesel fuel was aimed at reducing the pollution produced by diesel engines. Biodiesel fuel is recyclable and has low emission characteristics. One aspect of the study is the droplet size (μm) injected into the engine, at a fixed distance from the nozzle. Consider the following observed droplet size :

2.1 2.2 2.2 2.3 2.3 2.4 2.5 2.5 2.5 2.8 2.9 2.9 2.9 3.0 3.1 3.1 3.2 3.3 3.3 3.3 3.4 3.5 3.6 3.6 3.6 3.7 3.7 4.0 4.2 4.5 4.9 5.1 5.2 5.3 5.7 6.0 6.1 7.1 7.8 7.9 8.9

(*Source :* Kim et al (2008), Energy and Fuels, 22, 2091–2098.)

a) Group these droplet sizes and obtain a frequency table using $[2, 3)$, $[3, 4)$ and $[4, 5)$ as the first three classes, but try larger classes for the other cases.

b) Construct a density histogram.

c) Obtain the sample mean $\bar{x}$ and the sample variance s^2.

End of the tutorial class for Week 3.

MATH2099/MATH2859
Probability, Statistics and Information

Statistics Laboratory Class Week 4 : Descriptive Statistics

Instructions

After this tutorial you should be able to: draw density histograms; comment on density histograms; compute the sample mean, the sample variance, the sample standard deviation and the five-number summary of a data set; draw boxplots and side-by-side boxplots.

*Browse the general 'Introduction to Matlab' document and the 'Statistics Using Matlab' (here referred to as **SUM**) document to find extra information. Note that you can also use the* `help` *command at any time to get information about any* MATLAB *command.*

Do not just enter MATLAB *commands! Make sure you understand what you do and the output you obtain. When you produce a graph, look at it! What does it tell you about the data? Ask you tutor if you have questions. Explore the possibilities of Matlab. Go beyond what is actually required!*

Solutions to these exercises will be available from the course web page at a later date.

Exercise 1*

The data on inflows to the Hardap Dam (see Lab 2 Exercise 1) have been stored in the text file *inflows.txt*. Recover the vectors `Year` and `Inflow` which were imported from the *inflows.txt* data set into MATLAB in the first lab class.

a) Use MATLAB to find the mean, the variance, the standard deviation and the five-number summary of the inflows.

b) Construct a **density histogram** to display the Inflows data. Give the graph a title, and label the axes appropriately. The `histogram` function can construct a density histogram using the option `'Normalization', 'pdf'`. For example to plot a density histogram for data `x` with 5 classes you would type: `histogram(x,5,'Normalization','pdf')`

c) How does the number of classes changes the histogram? Plot density histograms with 5 and 50 classes and comment on how informative they are.

d) Equal-width classes may not be a good choice if a data set "stretches out" to one side or the other - as it is the case for the *inflows* data set. Add an `edges` vector to the `histogram` function to plot a density histogram with classes $[0, 500)$, $[500, 1500)$, $[1500, 3000)$ and $[3000, 6500)$.

e) Comment on the shape of the histogram. Without looking at the summary statistics, which one do you think is larger : the mean or the median? Why?

f) Construct a frequency histogram with the same classes as in c). Compare it to the density histogram obtained in c). Which one is misleading and why?

g) Write an **m.file** to carry out some or all of this exercise.

Exercise 2

The file *presdat2012.txt* contains data for the 44 US presidents and 27 Australian prime ministers to date giving the age at taking office. Recover the vectors `Name`, `Age` and `Country`, imported from the *presdat2012.txt* data set into MATLAB in the first lab class.

a) Calculate the sample mean, the sample variance and the sample standard deviation for ages at taking office in the two countries and comment. Remember the command `strcmp` to compare character string (see Lab 2, Exercise 2).

b) Construct side-by-side horizontal boxplots of age for the US and Australian leaders. Type `help boxplot` if you are unsure how to do this.

c) Calculate the sample median, the sample quartiles and the sample interquartile range for each country. Identify these on the boxplots.

d) The boxplot for the US shows two outlying values. Find the names of these two presidents who were significantly older than the others when they took office.

e) Do the boxplots suggest that there is a difference in ages at taking office for leaders in the two countries?

f) Write an **m.file** to carry out some or all of this exercise.

Exercise 3*

In a study designed to test the safety of various types of vehicles, vehicles were crashed into a wall at 60 km/h with a crash-test dummy strapped in the drivers seat. In the *crash.txt* data file, you will find the 6 variables

- `Head`: measure of head injuries
- `Chest`: chest deceleration (in g), which is a measure of chest injury
- `Airbag`: 1 indicates an airbag, 0 indicates only seatbelts were used
- `Doors`: number of doors (equals 6 for vans and sport utility vehicles)
- `Year`: year of vehicle
- `Weight`: weight of vehicle

Recover the vectors `Head`, `Chest`, `Airbag`, `Doors`, `Year` and `Weight` imported from the *crash.txt* data set into MATLAB from Moodle.

a) i) `Chest` is a measure of chest injuries sustained by the crash-test dummy. Larger values indicate more severe injuries. Use MATLAB to find the mean, the variance, the standard deviation and the five-number summary of the chest injuries.

ii) Determine the proportion of observations that have a value of `Chest` that is larger than or equal to 60.

iii) Construct a density histogram for `Chest`. How many classes would you consider? Comment on its shape. Experiment with changing the number of classes. For instance, produce a density histogram with 6 classes, and one with 25 classes. Compare and comment.

b) Suppose we wanted to see if airbags prevented injuries. Since `Airbag` is a categorical variable and `Chest` is a quantitative variable, a good form of analysis would be a side-by-side boxplot. Type `help boxplot` if you are unsure how to do this.

   i) Compare the distribution of chest deceleration for airbag (`Airbag = 1`) and non-airbag (`Airbag = 0`) cars. Do airbags appear to prevent injury?

   ii) How do the spreads of the two distributions compare?

   iii) Are there any outliers?

   iv) Is either distribution skewed? If so, describe the skewness.

c) Now lets compare the chest injuries of vehicles with different numbers of doors. Create side-by-side boxplots, as above, using these two variables.

   i) Does there appear to be a relationship between the number of doors and chest deceleration? What about the plot suggests a relationship?

   ii) Which type of vehicle (in terms of number of doors) tends to have the least severe injuries?

   iii) Why do you think that chest deceleration might be related to number of doors on the vehicle?

d) To test one reasonable theory, make a boxplot of `Weight` by `Doors`. What do these boxplots suggest about the results of the previous boxplot comparing chest deceleration by number of doors?

e) Write an **m.file** to carry out some or all of this exercise.

End of Laboratory class for Week 4.
If you are unable to finish during the class then try to finish in your own time before your next lab class so that you can seek help then if needed.

MATH2099/MATH2859
Probability, Statistics and Information

Statistics Tutorial Class Week 5 : Probabilities and Random Variables

Instructions

The following are exercises that will covered during your tutorial class.

Note that ***not every topic in the course can be covered by the tutorial exercises.*** *You are expected to take an active part in the learning process by working by yourself on most of the topics. In particular,* ***you are expected to attempt these tutorial questions beforehand****. During the tutorial, the tutor will go through the answers to some of the questions while directing explanation to areas where students indicate they have difficulty.*

Solutions to these exercises will be available from the course web page at a later date.

Exercise 1

For any collection of events $A_1, A_2, A_3, \ldots, A_k$, it can be shown that the inequality

$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup A_4 \cup \ldots \cup A_k) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_k)$$

always holds (try to understand why! *Hint:* use Venn diagrams and the additive law of probability). This inequality is most useful in cases where the events involved have relatively small probabilities. For example, suppose a system consists of five subcomponents connected in series and that each component has a 0.01 probability of failing. Find the upper bound on the probability that the entire system fails.

For any collection of events $A_1, A_2, A_3, \ldots, A_k$, it can be shown that the inequality

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4 \cap \ldots \cap A_k) \geq 1 - \big(\mathbb{P}(A_1^c) + \mathbb{P}(A_2^c) + \ldots + \mathbb{P}(A_k^c)\big)$$

always holds (try to understand why! *Hint:* use De Morgan's laws and the first part of this exercise). This inequality is most useful in cases where the events involved have relatively high probabilities. For example, suppose a system consists of ten subcomponents connected in series and that each component has a 0.999 probability of functioning without failure. Find the lower bound on the probability that the entire system functions correctly.

Exercise 2

Among the students doing a given course in engineering, there are four males enrolled in the civil engineering program, six females enrolled in the civil engineering program, and six males enrolled in the chemical engineering program. How many girls must be enrolled in the chemical engineering program if gender and engineering program are to be independent when a student is selected at random?

Exercise 3*

You are imprisoned in a dungeon together with two fellow prisoners. You are informed by the jailer that one of you has been chosen at random to be hanged, and the other two are to be freed. You ask the jailer to tell you privately which of your fellow prisoners will be set free, claiming that there would be no harm in divulging this information, since you already know that at least one of them will go free.

a) The jailer refuses to answer the question, pointing out that if you knew which of your fellows were to be set free, then your own probability of being executed would rise from 1/3 to 1/2, since you would then be one of two prisoners. Show that the jailer is wrong.

b) You convince the jailer, and he tells you which of the other two will be set free. You say this information to your fellow prisoners. While the spared guy is jumping for joy, the other one asks you to switch your identities. Would you accept?

Exercise 4*

An oil exploration company currently has two active projects, one in Asia and the other in Europe. Let A be the event that the Asian project is successful and B the event that the European project is successful. Suppose that A and B are independent events with $\mathbb{P}(A) = 0.4$ and $\mathbb{P}(B) = 0.7$.

a) If the Asian project is not successful, what is the probability that the European project is also not successful?
b) What is the probability that at least one of the two projects will be successful?
c) Given that at least one of the two projects is successful, what is the probability that only the Asian project is successful?

Exercise 5

Let X be a r.v. with cumulative distribution function $F(x)$ and density $f(x) = F'(x)$. Find the probability density function of

a) the maximum of n independent random variables all with cumulative distribution function $F(x)$.
b) the minimum of n independent random variables all with cumulative distribution function $F(x)$.

Exercise 6

An article in the review *Knee Surgery, Sports Traumatology and Arthroscopy* in 2005 cites a success rate of more than 90% for meniscal tears with a rim width of less than 3mm, but only a 67% success rate for tears of 3–6mm. If you are unlucky enough to suffer from a meniscal tear of less than 3mm on your left knee and one of width 3–6mm on your right knee, what is the probability mass function of the number of successful surgeries? Assume the surgeries are independent. Find the mean and variance of the number of successful surgeries that you would undergo.

Exercise 7*

The article "*Error Distribution in Navigation*" (J. Institute of Navigation, 1971) suggests that the distribution of the lateral position error, say X (in nautical miles), which can be either positive or negative, is well approximated by a density like

$$f(x) = c \times e^{-0.2\,|x|} \qquad \text{for } -\infty < x < \infty,$$

for a constant c.

a) Find the value of c which makes f a legitimate density function, and sketch the corresponding density curve.
b) In the long-run, what proportion of errors is negative? At most 2? Between -1 and 2?

Exercise 8

The probability density function of the weight X (in kg) of packages delivered by a post office is

$$f(x) = \frac{70}{69x^2} \qquad \text{for } 1 < x < 70$$

and 0 elsewhere.

a) Determine the mean and the variance of the weight X.
b) If the shipping cost is \$2.50 per kg, what is the average shipping cost of a package? What is the variance of the shipping cost?
c) In the long-term, what is the proportion of packages whose weight exceeds 50 kg?

Exercise 9*

a) Given the four scatter plots for two random variables X and Y in Figure 1,
   i) which of the plots demonstrates a positive relationship ?
   ii) which of the plots demonstrates a positive linear relationship ?
   iii) which of the plots demonstrates a negative relationship ?
   iv) which of the plots demonstrates no relationship ?

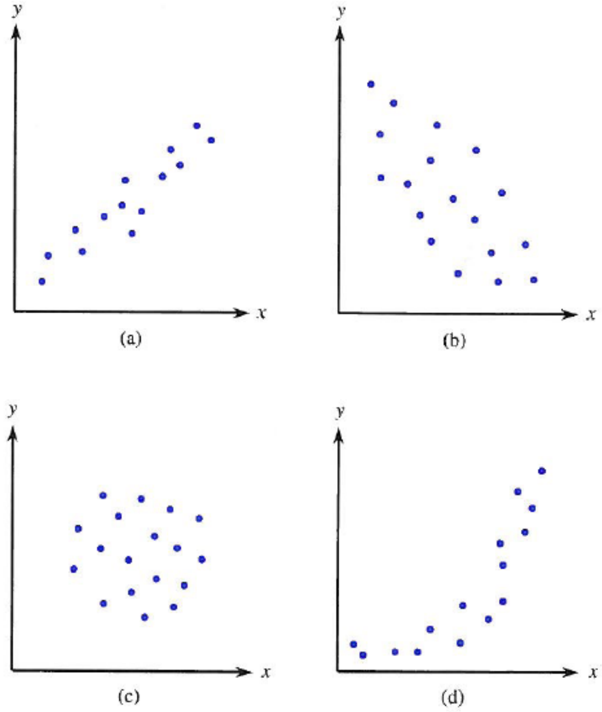


FIGURE 1

v) for which of the plots would you expect positive correlation ? Negative correlation ? No or little correlation ?

b) For each of the following pairs of variables, indicate whether you would expect a positive correlation, a negative correlation, or little or no correlation. Explain your choices.
   i) Maximum daily temperature and cooling cost
   ii) Interest rate and number of loan applications
   iii) Distance a student doing MATH2099 lives from UNSW campus and their marks at the Matlab online quizzes

Exercise 10

Many families of probability distributions are useful for engineers, including **Binomial** and **Poisson**.

a) In each of the following situations state whether it is reasonable to use one of these distributions for X. If so, tell which one, and (if it is possible) determine what are the values of the parameters :
   i) Toss a fair coin 6 times, X is the number of 'Heads'.
   ii) Toss a fair coin until the first time a head appears, X is the count of the number of tosses you make.
   iii) A factory makes carpets. Sometimes there are flaws in the carpet. On average a square metre of carpet has 3 flaws. X is the number of flaws in a random square metre of carpet.
   iv) Most calls made at random by sample surveys don't succeed in talking to a person. Of calls in New York City, only 1/12 succeed. A survey calls 500 randomly selected numbers in New York City, and X is the number that reach a live person.
   v) Calls to a telephone exchange come in at an average of 250 an hour, X is the number of calls in a given hour.
   vi) A die (6 faces, numbered 1,2,3,4,5,6) is tossed twice and X is the number of 6s obtained.
   vii) A die (6 faces, numbered 1,2,3,4,5,6) is tossed 7 times and X is the largest number obtained in the 7 tosses.
   viii) A die (6 faces, numbered 1,2,3,4,5,6) is tossed 4 times and X is the total of the numbers obtained on the top face.

b) Select one of the variables in part b) which can be modelled by a **Binomial distribution**, and find the long-run proportion of times that $X \leq 1$.

c) Select one of the variables in part b) which can be modelled by a **Poisson distribution**, and find the long-run proportion of times that $X \leq 1$.

d) In October 1994, a flaw in a certain Pentium chip installed in computers was discovered that could result in a wrong answer when performing division. The manufacturer initially claimed that the chance of any particular division being incorrect was only 1 in 9 billion, so that it would take thousands of years before a typical user encountered a mistake. However, statisticians are not typical users; some modern statistical techniques are so computationally intensive that a billion divisions over a short period of time in not outside the realm of possibility. Assuming that the 1 in 9 billion figure is correct and that results of divisions are independent of one another, what is the probability that at least one error occurs in one billion divisions with this chip?

Exercise 11

An individual claims to have extrasensory perception (ESP). As a test, a fair coin is tossed ten times, and he is asked to predict in advance the outcome. Our individual gets seven out of ten correct. What is the probability he would have done at least this well if he had no ESP? Would you believe in his powers?

Exercise 12

a) Suppose a value Z is repeatedly randomly chosen from a standard normal distribution:

   i) In the long run, what is the proportion of times that Z will be at most 2.15? Less than 2.15?

   ii) What is the long run proportion of times that Z will be between -1.23 and 2.85?

   iii) What is the long run proportion of times that Z will exceed 5? Will exceed -5?

   iv) What is the long run proportion of times that Z will satisfy $|Z| < 2.50$?

b) The article "Characterization of Room Temperature Damping in Aluminum-Indium Alloys" (*Metallurgical Trans.*, 1993) suggests that Al matrix grain size (μm) for an alloy consisting of 2% Indium could be modeled with a normal distribution with a mean value 96 and standard deviation 14.

   i) What is the probability that grain size exceeds 100?

   ii) What is the probability that grain size is between 50 and 80?

   iii) What interval (a, b) includes the central 90% of all grain sizes (so that 5% are below a and 5% are above b)?

End of the tutorial class for Week 5.

MATH2099/MATH2859
Probability, Statistics and Information

Statistics Laboratory Class Week 6 : Probability and Distributions

Instructions

After this tutorial you should be able to: use MATLAB *for determining probabilities from common probability distributions.*

*Browse the general 'Introduction to Matlab' document and the 'Statistics Using Matlab' (here referred to as **SUM**) document to find extra information. Remind that you can also use the* `help` *command at any time to get information about any* MATLAB *command.*

Do not just enter MATLAB *commands! Make sure you understand what you do and the output you obtain. When you produce a graph, look at it! What does it tell you about the data? Ask you tutor if you have questions. Explore the possibilities of Matlab, go beyond what is actually required!*

Solutions to these exercises will be available from the course web page at a later date.

MATLAB has pdf and cdf commands for determining probabilities from many distributions, including **Normal**, **Exponential**, **Uniform**, **Binomial** and **Poisson**. Use the `help` command for more details, or see Section 4 of the SUM. Note that MATLAB uses the notation "pdf" for both the probability mass function (pmf) of discrete distributions and the probability density function (pdf) of continuous distributions.

Exercise 1: Probability of events

Consider tossing a fair die. Define the events $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Then, $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 2/3$ and $\mathbb{P}(A \cap B) = 1/3$ (check this!). This shows that the events A and B are independent (why?). Using the MATLAB command `randi` (use `help` to determine how to use it), simulate draws from the sample space a large number of times (say, $N = 1,000$ times) and estimate the probabilities by the observed proportions p_A, p_B and p_{AB}. Verify that $p_A \times p_B \simeq p_{AB}$. Now find two events which are not independent. Repeat.

Exercise 2: the Binomial distribution*

The Binomial distribution is a discrete distribution. Use `help` to determine how to use `binopdf` and `binocdf`.

a) For $Y \sim Bin(20, 0.35)$, compute
   i) $\mathbb{P}(Y = 5)$
   ii) $\mathbb{P}(Y = 0)$
   iii) $\mathbb{P}(Y \leq 0)$, and compare ii)
   iv) $\mathbb{P}(5 \leq Y < 15)$
   v) $\mathbb{P}(5 < Y \leq 15)$
   vi) $\mathbb{P}(Y \in \{1, 5, 9, 17\})$
   vii) $\mathbb{P}(Y \notin \{5, 13, 16\})$

b) The multiple-choice Statistics Online Quiz 1 contained 10 questions, each with 5 answers. Assume a student just guessed on each question.
   i) What was the probability that the student answered all the questions correctly?
   ii) What was the probability that the student answered at least 5 questions correctly?

   Suppose the three Statistics Online Quizzes over the semester are similarly designed (10 questions, each with 5 answers)
   iii) What is the probability that the student, guessing on each question in each quiz, passes (at least 5 correct answers out of 10 questions) on at least one quiz?

Exercise 3: the Poisson distribution*

The Poisson distribution is a discrete distribution. Use `help` to determine how to use `poisspdf` and `poisscdf`.

a) For $Q \sim \mathcal{P}(7)$, compute
   i) $\mathbb{P}(Q \leq 3)$
   ii) $\mathbb{P}(Q < 3)$
   iii) $\mathbb{P}((Q = 19) \cup (Q = 20))$
   iv) $\mathbb{P}(Q > 0)$

b) Astronomers treat the number of stars in a given volume of space as a Poisson random variable. On average, in the Milky Way galaxy in the vicinity of our Solar System, they find one star per 16 cubic light-years.
   i) What is the probability of two or more stars in 16 cubic light-years?
   ii) How many cubic light-years of space must be studied so that the probability of one or more stars exceeds 0.95? (*Hint*: recall the Poisson process (Slide 34, Week 5))

Exercise 4: the Uniform distribution*

The Uniform distribution is a continuous distribution. Use `help` to determine how to use `unifpdf` and `unifcdf`. You may also find useful the command `unifinv` (use `help` to find out what it is for).

a) For $X \sim U_{[-1,1]}$, find
   i) $\mathbb{P}(X < 0)$
   ii) $\mathbb{P}(X \leq 0)$
   iii) $\mathbb{P}(-0.9 \leq X \leq 0.8)$
   iv) the value of x such that $\mathbb{P}(-x \leq X \leq x) = 0.9$

b) The thickness of photoresist applied to wafers in semiconductor manufacturing at a particular location on the wafer is uniformly distributed between 0.2050 and 0.2150.
   i) Determine the proportion of wafers that exceeds 0.2125 micrometers in photoresist thickness.
   ii) What thickness is exceeded by 10% of the wafers?

Exercise 5: the Exponential distribution*

The Exponential distribution is a continuous distribution. Use `help` to determine how to use `exppdf` and `expcdf`. You may also find useful the command `expinv` (use `help` to find out what it is for).

a) For $W \sim \text{Exp}(2)$, find
   i) $\mathbb{P}(W \leq 2)$
   ii) $\mathbb{P}(W < 2)$
   iii) $\mathbb{P}(10 < W < 13)$
   iv) $\mathbb{P}(W > -5)$

b) Suppose that the time to failure (in hours) of fans in a personal computer can be modeled by an exponential distribution with $\mu = 1/0.0003$.
   i) What proportion of fans will last at least 10,000 hours?
   ii) What proportion of fans will last at most 7000 hours?
   iii) 95% of the fans will last for at least which amount of time?

Exercise 6: the Normal distribution*

The Normal distribution is a continuous distribution. Use `help` to determine how to use `normpdf` and `normcdf`. You may also find useful the command `norminv` (use `help` to find out what it is for).

a) For $Z \sim \mathcal{N}(0, 1)$, compute
   i) $\mathbb{P}(-1 < Z < 1)$
   ii) $\mathbb{P}(-2 < Z < 2)$
   iii) $\mathbb{P}(-3 < Z < 3)$
   iv) Comment on i), ii) and iii) (look at Slide 11 Week 6)
   v) What is the value of z such that $\mathbb{P}(Z > z) = 0.05$?

vi) What is the value of z such that $\mathbb{P}(Z > z) = 0.025$?
vii) What is the value of z such that $\mathbb{P}(Z > z) = 0.005$?
viii) Comment on v), vi) and vii) (look at Slide 19 Week 6)

b) For $X \sim \mathcal{N}(3, 2)$, compute
i) $\mathbb{P}(2 < X < 4)$
ii) $\mathbb{P}(2 \leq X \leq 4)$
iii) $\mathbb{P}(X \geq 4)$
iv) $\mathbb{P}(1 < X < 5)$, and compare to a) i).

c) An article in the review *Archives of Environmental and Occupational Health* considered polycyclic aromatic hydrocarbons and immune system function in beef cattle. Some cattle were near major oil- and gas-producing areas of western Canada. The mean monthly exposure to PM1.0 (particulate matter that is $\leq 1\mu$m in diameter) was approximately 7.1 μg/m^3 with standard deviation 1.5 μg/m^3. Assume the monthly exposure is normally distributed.
i) What is the probability of a monthly exposure greater than 9 μg/m^3?
ii) What is the probability of a monthly exposure between 3 and 8 μg/m^3?
iii) What is the monthly exposure that is exceeded with probability 0.05?

Exercise 7: the disttool command in Matlab

a) Start the MATLAB demo on distributions by typing `disttool` in the Command Window. A graph showing the cumulative distribution function (CDF) for the standard normal distribution should appear. You can view the probability density function (PDF) by changing the Function type from CDF to PDF. To change the mean and standard deviation you can use the sliding bars labelled Mu (μ) and Sigma (σ), or type in values. Experiment with these features to see the effect of changing the mean and standard deviation.

b) What happens to the pdf as the mean is increased or decreased?

c) What happens to the pdf as the standard deviation is increased or decreased?

d) For the normal random variable $\mathcal{N}(3, 1)$, find the value of the cdf at $x = 2$. (*Hint*: change the mean and standard deviation, then type the value 2 in the X box, click anywhere in the grey area outside the graph to update it, and look at the value in the Probability box). Interpret this value in terms of a probability statement.

e) For the normal random variable $\mathcal{N}(3, 1)$ find the value of x that cuts off probability 0.05 in the upper tail (*Hint*: this will be 0.95 in the lower tail. Use the Probability box). Interpret this value in terms of a probability statement.

f) Experiment with the `disttool` tool with the other distributions introduced in class: Binomial, Poisson, Uniform and Exponential.

End of Laboratory class for Week 6.
If you are unable to finish during the class then try to finish in your own time before your next lab class so that you can seek help then if needed.

MATH2099/MATH2859
Probability, Statistics and Information

STATISTICS LABORATORY CLASS WEEK 8 :
THE CENTRAL LIMIT THEOREM AND CONFIDENCE INTERVALS

Instructions

After this tutorial you should be able to: use simulations in MATLAB *to explore the sampling distribution of the sample mean and to illustrate the Central Limit Theorem, use* MATLAB *to derive confidence intervals of a given level, and check if the normal assumption is reasonable.*

Browse the general 'Introduction to Matlab' document and the 'Statistics Using Matlab' (here referred to as ***SUM****) document to find extra information. Remind that you can also use the* `help` *command at any time to get information about any* MATLAB *command.*

Do not just enter MATLAB *commands! Make sure you understand what you do and the output you obtain. When you produce a graph, look at it! What does it tell you about the data? Ask your tutor if you have questions. Explore the possibilities of Matlab. Go beyond what is actually required!*

Solutions to these exercises will be available from the course web page at a later date.

In MATLAB you can simulate random samples from known distributions, using a `-rnd` command (`rnd` for 'random'). For instance, you can obtain a sample column vector of 5 values from the Normal distribution $\mathcal{N}(12, 3)$ by entering the command `normrnd(12,3,5,1)`. If you just enter `normrnd(12,3,5)`, then you will get a 5×5 matrix of random values from $\mathcal{N}(12, 3)$. To obtain samples from the Exponential distribution, use `exprnd`, to obtain samples from the Uniform distribution, use `unifrnd`, to obtain samples from the Binomial distribution, use `binornd`, to obtain samples from the Poisson distribution, use `poissrnd`. Use `help` for more information. See also Section 5 of the SUM document.

Exercise 1*

In this question we consider lots of simulated random samples of size 100 from a Binomial distribution.

a) Simulate (use the MATLAB command `binornd`) a matrix `B` with 100 rows and 500 columns containing elements that are independent values drawn from the Binomial distribution with $n = 1$ and $\pi = 0.5$ (that is, the Bernoulli distribution with parameter $\pi = 0.5$). You can think of each column of `B` as being a random sample of size 100 from $X \sim \text{Bern}(0.5)$. With 500 columns in `B`, we have 500 independent random samples of size 100.

b) Use the MATLAB `mean` function to compute the means of each of the 500 columns. You get a row vector, call it `meanB` (say), of length 500, which consists of the means of 500 samples of size 100, that is 500 values of the random variable $\bar{X} = \frac{1}{100}(X_1 + X_2 + \ldots + X_{100})$.
Note that, when applied to a matrix, the MATLAB function `mean` computes the mean of each column and returns a vector.

c) From lectures, what do you expect the mean and variance of the sampling distribution of the random variable $\bar{X}$ to be? Calculate the sample mean and variance for `meanB` and compare with your expectations.

d) From lectures, what do you expect the sampling distribution of the random variable $\bar{X}$ to be? Recall the Central Limit Theorem. Construct a density histogram of the simulated sample means. Comment on the shape.

e) The `histfit` function can overlay a histogram with the probability density function of any distribution. To plot a histogram of `meanB` with 10 classes, overlaid with the Normal distribution type `histfit(meanB,10,'normal')`, see `help histfit` for other distributions. Comment on how well the Normal distribution fits the data.

f) Write an **m.file** to carry out some or all of this exercise.

Exercise 2

The file *kevlar90.txt* contains times to failure (in hours) for a sample of size $n = 101$ of strands of Kevlar 49/epoxy, a material used in the space shuttle, when tested at a stress level of 90%. Recover the vector `kevlar90`, imported from the *kevlar90.txt* data set into MATLAB in the first lab class.

a) i) Construct a density histogram for the kevlar90 data. Comment on the shape. Do you think that a Normal distribution would fit the data well? Do you think that an Exponential distribution would fit the data well?

ii) Find the sample mean $\bar{x}$ of the 101 failure times.

iii) From i), we can assume that the Exponential distribution is the population distribution, and that the sample is a random sample drawn from it. Recall that the expectation of the Exponential(μ)-distribution is μ, which can thus be regarded as the population mean. We can estimate the 'true' Exponential distribution Exp(μ) by the Exponential distribution with parameter $\bar{x}$: we just replace in the distribution the unknown value of the mean (μ) by its natural estimate, the sample mean $\bar{x}$ ($\rightarrow \hat{\mu} = \bar{x}$). This simple way of fitting a distribution is known as the *method of moments*. Write down the fitted Exponential density for the Kevlar failure times. Given that the standard deviation of the Exp(μ) density is also μ, state the standard deviation of a random variable with the fitted Exponential density. Compare the sample standard deviation of the data set `kevlar90`.

iv) Use the `histfit` function to overlay the histogram of the Kevlar data with the fitted Exponential density and briefly comment on how well the exponential density fits the data.

v) Write an **m.file** to carry out some or all of this exercise.

b) Next we examine the sampling distribution of the mean failure time for a sample of 10 strands of Kevlar.

i) Use the MATLAB command `exprnd` to simulate a matrix `T` with 10 rows and 1000 columns containing independent elements that are Exponentially distributed with mean equal to the sample mean of the Kevlar failure times. We can think of each column of `T` as being failure times for a sample of 10 strands of Kevlar similar to those in the dataset. With 1000 columns in `T`, we have 1000 random samples of size 10. Each of them could have been the observed sample in *kevlar90.txt*! It must be clear that here the population distribution which generated the samples is $\mathrm{Exp}(\hat{\mu})$, with $\hat{\mu}$ found in a)iii).

ii) Use the `mean` function to compute the means of the 1000 columns. Thus, you get a row vector, called `meanT` (say), of length 1000, which consists of the means of 1000 samples of size 10, that is 1000 values of the random variable $\bar{X} = \frac{1}{10}(X_1 + X_2 + \ldots + X_{10})$.

iii) From lectures, what do you expect the mean and variance of the sampling distribution of the random variable $\bar{X}$ to be? Calculate the sample mean and variance for `meanT` and compare with your expectations.

iv) From lectures, what do you expect the sampling distribution of the random variable $\bar{X}$ to be? Recall the Central Limit Theorem. Construct a density histogram of the simulated sample means. Comment on the shape.

v) Use the `histfit` function to overlay the expected large sample density on the histogram and briefly comment on the level of similarity.

vi) Write an **m.file** to carry out some or all of this exercise.

c) Repeat part b) but this time generating 1000 random samples of size $n = 500$.

Exercise 3*

The file *shearstrength.txt* contains 100 observations on shear strength (in lb) of ultrasonic spot welds made on a certain type of alclad sheet. This data is in agreement with the one in *Comparison of Properties of Joints Prepared by Ultrasonic Welding of Other Means*, Journal of Aircraft, 1983: 552-556. Recover the vector `shearstrength`, imported from the *shearstrength.txt* data set into MATLAB in the first lab class.

a) Construct a density histogram of the shear strengths. Comment on the shape of the distribution.

b) Is it plausible that the sample was selected from a normal distribution? Draw a normal quantile plot and conclude. Note: in MATLAB, a normal quantile plot is represented by entering the command `qqplot`. Use `help qqplot` for more information.

c) Investigators believe that the population standard deviation of shear strength in this context is $\sigma = 350$ lb. Estimate μ, the true mean shear strength and give a 95% two-sided confidence interval for it. Interpret your result. Note: in MATLAB, this type of 'z-confidence interval' (see Slide 19, Week 8 of the lecture slides) is obtained with the command `ztest`. This is actually a command used for hypothesis testing (which we have not yet studied), but is also used for determining confidence intervals. For instance, to calculate the 95% CI for a population mean from a sample contained in the vector `data`, use the command

$$[\mathtt{h}, \mathtt{p}, \mathtt{ci}] = \mathtt{ztest(data, xbar, sigma); ci}$$

where `xbar` is the sample mean and `sigma` is the supposedly known population standard deviation σ. The default option is for a 95% confidence interval. If you require any other level of significance you need to enter a fourth parameter `alpha`. The outputs `h` and `p` refer to the underlying hypothesis test and are not relevant for us at this point; that is why we focus only on the output `ci` here. Type `help ztest` for more information. See also the SUM document, Section 6.1.

d) Now, suppose σ is not known. Give a 95% two-sided confidence interval for μ. Interpret your result. Compare with the previous confidence interval, and explain why they are different. Note: in MATLAB, this type of 't-confidence interval' (see Slide 19, Week 8 of the lecture slides) is obtained with the command `ttest`. Again, this is actually a command used for hypothesis testing but it is also used for determining confidence intervals. For instance, to calculate the 95% CI for a population mean from a sample contained in the vector `data`, use the command

$$[\mathtt{h}, \mathtt{p}, \mathtt{ci}] = \mathtt{ttest(data, xbar); ci}$$

where `xbar` is the sample mean. The default option is for a 95% confidence interval. If you require any other level of significance you need to enter a third parameter `alpha`. Type `help ttest` for more information. See also the SUM document, Section 6.1.

e) Write an **m.file** to carry out some or all of this exercise.

Exercise 4

In this exercise, we will use simulations to drive home the proper interpretation of confidence intervals.

a) Simulate (use the MATLAB command `normrnd`) a matrix `C` with 36 rows and 500 columns containing elements that are independent values drawn from the Normal distribution with $\mu = 20$ and $\sigma = 5$. You can think of each column of `C` as being a random sample of size 36 from $X \sim \mathcal{N}(20, 5)$. With 500 columns in `C`, we have 500 independent random samples of size 36.

b) Calculate 95% z- and t-confidence intervals for μ from each sample. Recall that the z-confidence interval of level $100 \times (1 - \alpha)\%$ is given by

$$\left[\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$$

where σ is assumed to be known, and the t-confidence interval of level $100 \times (1 - \alpha)\%$ is

$$\left[\bar{x} \pm t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}}\right]$$

where s is the sample standard deviation.

i) Use the MATLAB `mean` function to compute the means of each of the 500 samples. You get a row vector, call it `meanC` (say), of length 500, which consists of the 500 sample means.
Note that, when applied to a matrix, the MATLAB function `mean` computes the mean of each column and returns a vector.

ii) Use the MATLAB `std` function to compute the standard deviations of each of the 500 samples. You get a row vector, call it `sC` (say), of length 500, which consists of the 500 sample standard deviations.
(Like `mean`, when applied to a matrix, `std` computes the standard deviation of each column and returns a vector.)

iii) From the vector `meanC`, build the vector `upz` of upper bounds for the z-CI for each sample. Recall that $z_{0.975} = 1.96$ and that here, $n = 36$ and $\sigma = 5$ (if assumed to be known).

iv) From the vector `meanC`, build the vector `lowz` of lower bounds for the z-CI for each sample.

v) From the vectors `meanC` and `sC`, build the vector `upt` of upper bounds for the t-CI for each sample. Find the value $t_{35;0.975}$ in MATLAB (hint: use the `tinv` function).

vi) From the vectors `meanC` and `sC`, build the vector `lowt` of lower bounds for the t-CI for each sample.

c) You have now 500 z-confidence intervals and 500 t-confidence intervals for μ computed from 500 independent samples. Compute how many of the 500 z-confidence intervals and how many of the 500 t-confidence intervals contain the true population mean $\mu = 20$. Comment.

d) If you were to repeat this exercise over and over, on average, what fraction of the confidence intervals you calculate would you expect to contain the population mean?

e) Write an **m.file** to carry out some or all of this exercise.

Exercise 5

For each of 18 preserved cores from oil-west carbonate reservoirs, the amount of residual gas saturation after a solvent injection was measured at water flood-out. The data can be found in the file *porevolume.txt*, and are from "*Relative Permeability Studies of Gas Water Flow Following Solvent Injection in Carbonate Rocks*", Soc. Petroleum Engineers J., 1976: 23-30. Recover the vector `porevolume`, imported from the *porevolume.txt* data set into MATLAB in the first lab class.

a) Is it plausible that the sample was selected from a normal population distribution? Show a density histogram and a normal quantile plot and conclude.

b) Determine a 98% CI for the true average amount of residual gas saturation.

c) Determine a 95% CI for the true average amount of residual gas saturation. What is the difference between a 95% CI and a 98% CI?

d) Write an **m.file** to carry out some or all of this exercise.

Exercise 6

The file *fusiondat.txt* contains results from an experiment in visual perception using random dot sterograms. Two images appear to be composed entirely of random dots, but the images will fuse and a 3D object will appear when they are viewed in a certain way. An experiment was performed to determine whether knowledge of the form of the embedded image affected the time required for subjects to fuse the images. One group of subjects received either no information or just verbal information about the shape of the embedded object. A second group received both verbal information and visual information (e.g., a drawing of the object). Recover the matrix `fusiondat`, imported from the *fusiondat.txt* data set into MATLAB in the first lab class. Notice the form of the data file with the two columns. The variables in the file are the group the subject belongs to (1 = no information or just verbal information, 2 = verbal information and visual information), and the log of the time taken to fuse the images.

a) Which formula do you think is appropriate for constructing a confidence interval for the difference in mean log times for the two groups? Why?

b) Use this formula to construct a 95% confidence interval for the difference in mean log times for the two groups. Does the provision of visual information appear to make a difference in time taken to fuse the images?

c) Write an **m.file** to carry out some or all of this exercise.

Exercise 7

The file *abs.txt* contains data on internal deadband impedances for front and rear sensors from an antilock braking system (ABS). Each row of the dataset corresponds to the same unit, with the first measurement being the front sensor impedance and the second the rear sensor impedance (in kohms). Recover the matrix `abs`, imported from the *abs.txt* data set into MATLAB in the first lab class.

a) Which formula do you think is appropriate for constructing a confidence interval for the difference in mean impedance for the front and rear sensors? Why?

b) Construct a 95% confidence interval for the difference in mean impedance for front and rear sensors. Does location of sensor appear to affect impedance?

End of Computer Tutorial for Week 8.
If you are unable to finish during the class then try to finish in your own time before the lab test so that you can seek help before it if needed.

MATH2099/MATH2859
Probability, Statistics and Information

Statistics Tutorial Class Week 9 :
Estimation, Central Limit Theorem and Inferences concerning a mean

Instructions

The exercises that will be gone over during the tutorial class will be chosen from those printed here.

Note that ***not every topic in the course can be covered by the tutorial exercises.*** *You are expected to take an active part in the learning process, by working by yourself on most of the topics.*

In particular, ***you are expected to attempt these tutorial questions beforehand.*** *At the tutorial, the tutor will go through the answers to some of the questions, directing explanation to areas where students indicate they have difficulty.*

Solutions to these exercises will be available from the course web page at a later date.

Exercise 1*

a) Suppose that X is a random variable which is Uniform on the interval $[2, 5]$.
   i) Write down the density function of X. What are μ, σ^2 and σ for this distribution?
   ii) Random samples of size $n = 15$ are taken from the distribution of X. Determine the mean, variance and standard deviation of the sampling distribution of the sample mean $\bar{X}$.

b) Suppose that X is a random variable which is Exponentially distributed with mean 2.
   i) Write down the density function of X. What are μ, σ^2 and σ for this distribution?
   ii) Random samples of size $n = 12$ are taken from the distribution of X. Determine the mean, variance and standard deviation of the sampling distribution of the sample mean $\bar{X}$.

c) Suppose that X is a random variable which is Poisson distributed with mean 5.
   i) Write down the probability mass function of X. What are μ, σ^2 and σ for this distribution?
   ii) Random samples of size $n = 10$ are taken from the distribution of X. Determine the mean, variance and standard deviation of the sampling distribution of the sample mean $\bar{X}$.

Exercise 2

a) Suppose we have a random sample $X_1, X_2, \ldots, X_{2n}$ of size $2n$ from a population denoted by X, and $\mathbb{E}(X) = \mu$ and $\mathbb{V}\text{ar}(X) = \sigma^2$. Let

$$\bar{X}_1 = \frac{1}{2n}\sum_{i=1}^{2n} X_i \qquad \text{and} \qquad \bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

be two estimators of μ. Which is the better estimator of μ? Explain.

b) Let $X_1, X_2, \ldots, X_n$ denote a random sample from a population having mean μ and variance σ^2. Consider the following estimators of μ:

$$\hat{\mu}_1 = \frac{X_1 + X_2 + \ldots + X_7}{7}$$

$$\hat{\mu}_2 = \frac{2X_1 - X_6 + X_4}{2}$$

   i) Is either estimator unbiased?
   ii) Which estimator is best? In what sense is it best?

c) Two different plasma etchers in a semiconductor factory have the same mean etch rate μ. However, machine 1 is newer than machine 2 and consequently has smaller variability in etch rate. We know that the variance of etch rate for machine 1 is σ_1^2 and for machine 2 is $\sigma_2^2 = a\sigma_1^2$ $(a > 1)$. Suppose that we have n_1 independent observations on etch rate from machine 1 and n_2 independent observations on etch rate from machine 2.

   i) Show that $\hat{\mu} = \alpha\bar{X}_1 + (1-\alpha)\bar{X}_2$ is an unbiased estimator for μ for any value of α between 0 and 1.

   ii) Find the standard error of the point estimate of μ in part i).

   iii) What value of α would minimise the standard error of the point estimate of μ?

   iv) Suppose that $a = 4$ and $n_1 = 2n_2$. What value of α would you select to minimise the standard error of the point estimate of μ? How "bad" would it be to arbitrarily choose $\alpha = 0.5$ in this case? How "bad" would it be to only use the observations coming from machine 1?

Exercise 3*

a) A synthetic fibre used in manufacturing carpet has tensile strength that is normally distributed with mean 75.5 psi and standard deviation 3.5 psi. Find the probability that a random sample of $n = 6$ fibre specimens will have a sample mean tensile strength that exceeds 75.75 psi. How does this probability change when the sample size is increased from $n = 6$ to $n = 49$?

b) The compressive strength of concrete is normally distributed with $\mu = 2500$ psi and $\sigma = 50$ psi. Find the probability that a random sample of $n = 5$ specimens will have a sample mean strength that falls in the interval from 2499 to 2510 psi.

c) The amount of time that a customer spends waiting at an airport check-in counter is a random variable with mean 8.2 minutes and standard deviation 1.5 minutes. Suppose that a random sample of $n = 49$ customers is observed. Find the probability that the average waiting time for these customers is

   i) less than 10 minutes.

   ii) between 7 and 10 minutes.

   iii) less than 8.5 minutes.

Exercise 4

The Rockwell hardness of certain metal pins is known to have a mean of 50 and a standard deviation of 1.5.

a) Suppose that it is known that the distribution of all such pin hardness measurements is normal, then what can be said about the sampling distribution of the sample means from random samples of size n? Under this assumption of normality, what is the probability that the average hardness for a random sample of size $n = 9$ is at least 52?

b) What is the probability that the average hardness in a random sample of 40 pins is at least 52? Do you need to assume normality here?

Exercise 5*

The number of flaws X on an electroplated car grill is known to have the following probability mass function:

x	0	1	2	3
$p(x)$	0.8	0.1	0.05	0.05

a) Calculate the mean and standard deviation of X.

b) What are the mean and the standard deviation of the sampling distribution of the average number of flaws per grill in a random sample of 64 grills?

c) For a random sample of 64 grills, calculate (approximately) the probability that the average number of flaws per grill exceeds 0.5.

Exercise 6

An article in *Journal of Agricultural Science* (1997) investigated means of wheat grain crude protein content (CP) and Hagberg falling number (HFN) surveyed in the UK. The analysis used a variety of nitrogen fertiliser applications (kg N/ha), temperature (°C), and total monthly rainfall (mm). The data shown below describe temperatures for wheat grown at Harper Adams Agricultural College between 1982 and 1993. The temperatures measured in June were obtained as follows:

```
15.2 14.2 14.0 12.2 14.4 12.5 14.3 14.2 13.5 11.8 15.2
```

Assume that the standard deviation is historically known to be $\sigma = 0.5$ and that the temperature distribution is normal.

a) Construct a 99% two-sided confidence interval on the mean temperature.

b) Construct a 95% lower-confidence bound on the mean temperature (that is, a one-sided confidence interval).

c) Suppose that we wanted to be 95% confident that the error in estimating the mean temperature is less than 0.2°C. What sample size should be used?

Exercise 7

The wall thickness of 25 glass 2-litre bottles was measured by a quality-control engineer. The sample mean was $\bar{x} = 4.05$ millimetres, and the sample standard deviation was $s = 0.08$ millimetre. Find a 95% lower confidence bound for mean wall thickness. Interpret the interval you have obtained. Assume the normal distribution for wall thickness.

Exercise 8

While performing a certain task under simulated weightlessness, the pulse rate of 42 astronaut trainees increased by an average of 26.4 beats per minute with a standard deviation of 4.28 beats per minute.

a) Construct a two-sided 95% confidence interval for the true average increase in the pulse rate of astronaut trainees performing the given task.

b) What can one assert with 95% confidence about the maximum error if $\bar{x} = 26.4$ is used as a point estimate of the true average increase in the pulse rate?

Exercise 9*

Consider the following sample of fat content (in percentage) of $n = 10$ randomly selected hot dog sausages of a given brand ("Sensory and Mechanical Assessment of the Quality of Frankfurters", *Journal of Texture Studies*, 1990):

```
25.2 21.3 22.8 17.0 29.8 21.0 25.5 16.0 20.9 19.5
```

Assume that the fat content follows a normal population.

a) Find a 95% confidence interval for the true mean fat content of the sausages of that brand.

b) Suppose, however, you are going to eat a single hot dog of this type and want a prediction for the resulting fat content. Find a 95% prediction interval for the fat content of the hot dog sausage you will eat.

Exercise 10

The article "Repeatability and Reproducibility for Pass/Fail data" (*J. of Testing and Eval.*, 1997, 151-153) reported that in $n = 48$ trials in a particular laboratory, 16 resulted in ignition of a particular type of substrate by a lighted cigarette. Find a 95% approximate confidence interval for π, the true long-run proportion (or probability) of all such trials that would result in ignition.

End of the tutorial class for Week 9.

MATH2099/MATH2859
Probability, Statistics and Information

Statistics Tutorial Class Week 11 : Confidence Intervals and Hypotheses Tests

Instructions

The exercises that will be covered during the tutorial class will be chosen from those printed here.

Note that ***not every topic in the course can be covered by the tutorial exercises.*** *You are expected to take an active part in the learning process, by working by yourself on most of the topics.*

In particular, ***you are expected to attempt these tutorial questions beforehand****. During the tutorial, the tutor will go through the answers to some of the questions, directing explanation to areas where students indicate they have difficulty.*

Solutions to these exercises will be available from the course web page at a later date.

Exercise 1*

a) In a hypothesis test of H_0 with alternative H_a, the p-value is calculated to be 0.008. What would be the conclusion of the hypothesis test?

b) In a hypothesis test of H_0 with alternative H_a, the p-value is calculated to be 0.08. What would be the conclusion of the hypothesis test?

c) In a hypothesis test of H_0 with alternative H_a, the p-value is calculated to be 0.8. What would be the conclusion of the hypothesis test?

Exercise 2*

In the book 'Statistical quality design and control: Contemporary concepts and methods' DeVore, Chang, and Sutherland (1992) discuss a cyclinder boring process for an engine block. Specifications require that these bores be 3.5199 ± 0.0004 inches. Management is concerned that the true proportion of cylinder bores outside the specifications is excessive. Current practice is willing to tolerate up to 10% outside the specifications. Out of a random sample of 165 observations, 36 were outside the specifications.

a) Conduct the most appropriate hypothesis test using a 0.01 significance level.

b) Construct a 99% confidence interval for the true proportion of bores outside the specification.

c) What assumptions did you need to make to carry out this hypothesis test, and determine the confidence interval? Can you check these assumptions? If so, how?

Exercise 3*

The paper "Selection of a Method to Determine Residual Chlorine in Sewage Effluents" (Water and Sewage Works, 1971, pp. 360-364) reports the results of an experiment in which two different methods for determining chlorine content were used on specimens of Cl_2-demand-free water for various doses and contact times. Observations are in mg/l.

Specimen	1	2	3	4	5	6	7	8
MSI method	0.39	0.84	1.76	3.35	4.69	7.70	10.5	10.92
SIB method	0.36	1.35	2.56	3.92	5.35	8.33	10.70	10.91

The sample mean of the differences (MSI-SIB) between readings for these 8 samples is -0.4137 and the corresponding sample standard deviation is 0.3210.

a) Construct a 99% confidence interval for the true average of the difference of residual chlorine readings between the two methods (assume normality for the difference distribution).

b) Test whether the mean difference is zero at the 5% level.

c) What assumptions did you need to make to carry out this hypothesis test and to determine the confidence interval? Can you check any or all of these assumptions? If so, how?

Exercise 4

Let μ_1 and μ_2 denote true average tread lives for two competing brands of size P205/65R15 radial tyres. From samples of the two brands we have the data:

$$n_1 = 45,\ \bar{x}_1 = 42,500,\ s_1 = 2200,\ n_2 = 45,\ \bar{x}_2 = 40,400 \text{ and } s_2 = 1500.$$

a) Test $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$ at the $\alpha = 0.05$ level.

b) Compute a 95% confidence interval for $\mu_1 - \mu_2$. Does this interval suggest that $\mu_1 - \mu_2$ has been precisely estimated?

c) What assumptions did you need to make to carry out this hypothesis test and to determine the confidence interval? Can you check any or all of these assumptions? If so, how?

Exercise 5*

Are male university students more easily bored than their female counterparts? The following are data from a study in which a scale, called the *Boredom Proneness Scale*, was administered to 97 male and 148 female students. Does the following data support the research hypothesis that the mean Boredom Proneness Rating is higher for men than it is for women? Test the appropriate hypothesis at the 5% level of significance.

Gender	sample size	sample mean	sample sd
Male	97	10.40	4.83
Female	148	9.26	4.68

Exercise 6

Suppose μ_1 and μ_2 are the mean stopping distances (in m) at 80 km/h for cars of a certain type equipped with two different types of braking systems. An expert claims that, on average, cars equipped with braking system 2 need more than 10 m more to stop than cars equipped with braking system 1. Assuming normality of the stopping distances, use a two-sample t-test at a 1% level of significance to test $H_0 : \mu_1 - \mu_2 = -10$ against $H_a : \mu_1 - \mu_2 < -10$. The sample data yield:

$$n_1 = 6,\ \bar{x_1} = 115.7,\ s_1 = 5.03,\ n_2 = 6,\ \bar{x_2} = 129.3 \text{ and } s_2 = 5.38.$$

Exercise 7

In the article '*Estimating the current mean of a process subject to abrupt changes*' (Technometrics, 37, 311-323), Yashchin (1995) discusses a process for the chemical etching of silicon wafers used in integrating circuits. This company wishes to detect an increase in the thickness of the silicon oxide layers because thicker layers require longer etching times. Process specifications state a target value of 1 micron for the true mean thickness. Historically, the layer thickness has a standard deviation of 0.06 microns. You may assume that layer thickness is normally distributed.

a) A recent random sample of four wafers yielded a sample mean of $\bar{x} = 1.134$ microns. Conduct a hypothesis test to determine whether the true mean thickness has increased. Use a significance level of 0.05.

b) Using the sample information in a), construct a 95% confidence interval for the true current mean μ. Use this interval to determine whether the mean thickness has changed. Discuss the relationship of the 95% confidence interval and the corresponding hypothesis test.

c) Recall that the power of a test is an expression of its ability to detect when an alternative hypothesis is true (Slide 10, Week 9). The power is a function of the particular value considered under the alternative hypothesis, and is given by

$$\text{power} = 1 - \beta = \mathbb{P}(\text{reject } H_0 \text{ when it is false}).$$

Find the power of the test you used in part a) to detect a change in the true mean thickness to 1.01 (i.e if in fact $\mu = 1.01$). Interpret this value.

d) Find the sample size required to achieve a power of 0.85 when $\mu = 1.01$.

e) In this question you assumed that layer thickness was normally distributed. How could you check whether this is a plausible assumption? What else did you assume to do these analyses? Can you check these assumptions?

End of the tutorial class for Week 11.

MATH2099/MATH2859
Probability, Statistics and Information

Statistics Tutorial Class Week 12 : Regression Analysis

Instructions

The exercises that will be covered during the tutorial class will be chosen from those printed here.

Note that ***not every topic in the course can be covered by the tutorial exercises.*** *You are expected to take an active part in the learning process, by working by yourself on most of the topics.*

In particular, ***you are expected to attempt these tutorial questions beforehand****. During the tutorial, the tutor will go through the answers to some of the questions, directing explanation to areas where students indicate they have difficulty.*

Solutions to these exercises will be available from the course web page at a later date.

Exercise 1

In a certain chemical process the reaction time Y (in hours) is known to be related to the temperature X (in°F) in the chamber in which the reaction time takes place according to the simple linear equation

$$Y = 5.00 - 0.01X + \varepsilon,$$

where ε is a random disturbance normally distributed with standard deviation $\sigma = 0.075$.

a) What is the true average change in reaction time associated with a 1°F increase in temperature? A 10°F increase in temperature?

b) What is the true average reaction time when the temperature is 200°F? When the temperature is 250°F?

c) What is $\mathbb{P}(2.4 < Y < 2.6)$ when $X = 250$?

d) If an investigator makes five independent experimental runs for a temperature of 250°F what is the probability that all five observed reaction times are between 2.4 and 2.6 hours?

Exercise 2

A regression of Y = calcium content (g/l) on X = dissolved material (mg/cm^2) was reported in a 1997 article in the Magazine of Concrete Research. The equation of the fitted least squares regression line was determined :

$$\hat{y}(x) = 3.678 + 0.144 \times x \quad \text{with } r^2 = 0.860 \quad \text{and } n = 23.$$

a) Interpret the estimated slope 0.144 and the coefficient of determination 0.860.

b) What would be a point estimate of the true average calcium content when the amount of dissolved material is 50 mg/cm^2?

c) The value of the total sum of squares was $ss_t = 320.398$. Calculate an estimate of the error standard deviation in the simple linear regression model.

Exercise 3*

Mist (airborne droplets or aerosols) is generated when metal-removing fluids are used in machining operations to cool and lubricate the tool and work-piece. Mist generation is a concern of the OSHA (Occupational Safety and Health Administration), which has recently lowered the allowable mist generation standard for the workplace by a substantial amount. In a 2002 article in the Lubrication Engineering journal the following data was provided on :

X = fluid flow velocity for a 5% soluble oil (cm/sec)
Y = the extent of mist droplets having diameters smaller than 10 mm (mg/m^3)

x_i	89	177	189	354	362	442	965
y_i	0.40	0.60	0.48	0.66	0.61	0.69	0.99

Various regression output and graphics are included for this question. Use these in your answers to the following questions :

a) The investigators performed a simple linear regression analysis to relate the two variables. Does a scatter plot of the data support this strategy?

b) What proportion of observed variation in mist can be attributed to the linear relationship between velocity and mist?

c) The investigators were particularly interested in the impact on mist of increasing velocity from 100 to 1000 (a factor of 10 corresponding to the difference between the smallest and largest x-values in the sample). When X increases in this way is there substantial evidence (with $\alpha = 0.05$) that the true average increase in Y is less than 0.6?

d) Estimate the true average change in mist associated with a 1 cm/sec increase in velocity and do so in a way that conveys information about precision and reliability.

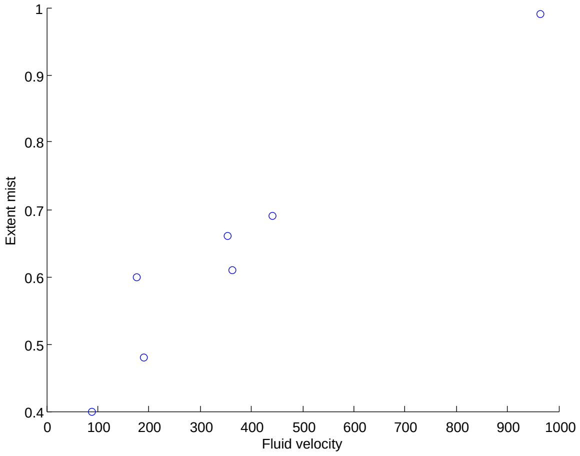
e) Comment on the normal quantile plot for residuals.

f) Provide a 95% confidence interval for the true average mist when fluid flow velocity is set at 500 cm/sec. Compare this with the graphical display.

g) Provide a 95% prediction interval for a future value of average mist that will be observed when the fluid flow is set at 500 cm/sec. Compare this with the graphic display.

h) Explain why the interval in part g) is wider than that in part f).

i) OHSA's draft standard requires that the mist exceed 1 mg/m^3 at most 2.5% of the time. Use the attached graph to determine the appropriate level of fluid velocity to meet this requirement.



```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                    Estimate          SE         tStat       pValue
                   __________    __________    ______    __________

    (Intercept)       0.40412      0.034589    11.684     8.071e-05
    x1             0.00062108    7.5792e-05    8.1945    0.00044032


Number of observations: 7, Error degrees of freedom: 5
Root Mean Squared Error: 0.054
R-squared: 0.931,  Adjusted R-Squared 0.917
F-statistic vs. constant model: 67.2, p-value = 0.00044
```

```
#predicted mean and confidence intervals
>> [ypred,yci] = predict(MyFit,500)
ypred =
    0.7147
yci =
    0.6562    0.7731

#prediction interval
>> [ypred,yci] = predict(MyFit,500,'Prediction','observation') ;yci
yci =
    0.5639    0.8654
```

That's it for tutorial exercises. You should try to do the rest in MATLAB on your own time.

Exercise 4

The file *rain.txt* contains rainfall (first column, in mm) and runoff (second column, in mm/h) measurements at Pontelagoscuro on the Po river in northeast Italy, for the 31 years 1918 and 1948. The data are to be used to produce a model for predicting runoff in terms of rainfall.

a) Recover the matrix `rain`, imported from the *rain.txt* data set into MATLAB in the first lab class. Define two vectors `rainfall` and `runoff` as the two columns of the matrix `rain`.

b) Produce a scatter-plot of runoff versus rainfall, and comment on the shape of the relationship between the two variables. Label appropriately. Write an equation for the regression model that you would like to fit to these data.

c) Using the MATLAB command `lsline`, add the least squares regression line to the plot. Enter `help lsline` if you need more information on this command.

d) Use the MATLAB command `fitlm` to fit a linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

to the data. Here, the response Y is the runoff and the predictor X is the rainfall.

Note: the `fitlm` function produces an object which contains just about everything you would want to know about the model. Some of the information is printed by default, while for others we have to use functions with the model object to extract information. For this question start by fitting the model and creating the linear model object using the following code:

```
>> rainMod=fitlm(rainfall,runoff)
```

From the produced output, answer the following questions:

(i) What is the equation of the fitted line?

(ii) What is the estimated value of σ, the standard deviation of the error term ε?

(iii) Can you conclude that the rainfall amount has a significant influence on the runoff? Test the relevant hypothesis at level $\alpha = 0.05$.

(iv) What is the estimated expected change in the runoff for a change in rainfall of 1 mm?

(v) Use the `coefCI` function to determine a 95% two-sided confidence interval for β_1.

(vi) What proportion of variation in the observed runoff is explained by the variation in the rainfall?

(vii) What is the (sample) correlation between rainfall and runoff?

(viii) Use the function `plotResiduals` to obtain a plot of the residuals versus the fitted values, and a normal quantile plot of the residuals, and comment briefly on these plots. Does that validate the model?

Exercise 5

Milk is an important source of protein. How does the amount of protein in milk from a cow vary with milk production? The article Metabolites of Nucleic Acids in Bovine Milk (Journal of Diary Science, 1984, 723-728) reported the data contained in the file *milk.txt* for Holstein Friesian cows.

a) Recover the matrix `milk`, imported from the *milk.txt* data set into MATLAB in the first lab class. The first column gives the milk production (in kg/day) for a cow and the second column the corresponding milk protein production (in kg/day). Define two vectors `x` and `y` as the two columns of the matrix `milk`.

b) Produce a scatter-plot of protein content (`y`) versus milk production (`x`), and comment on the shape of the relationship between the two variables. Label appropriately. Write an equation for the regression model that you would like to fit to these data.

c) Using the MATLAB command `lsline`, add the least squares regression line to the plot. Enter `help lsline` if you need more information on this command.

d) Use the MATLAB command `fitlm` to fit a linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

to the data. Here, the response Y is the protein content and the predictor X is the milk production.

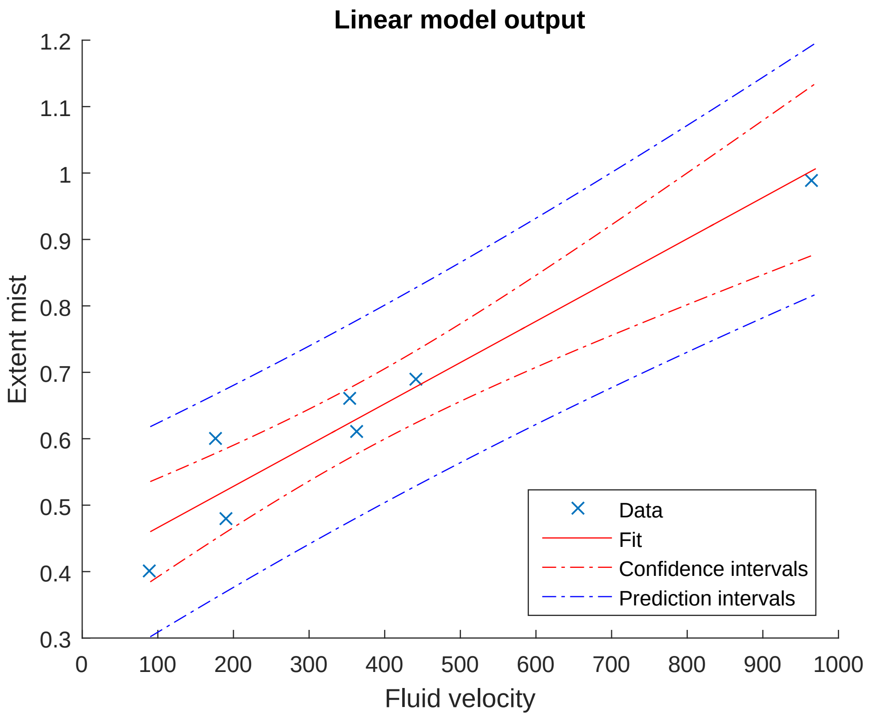
e) Write down the equation of the fitted regression line.

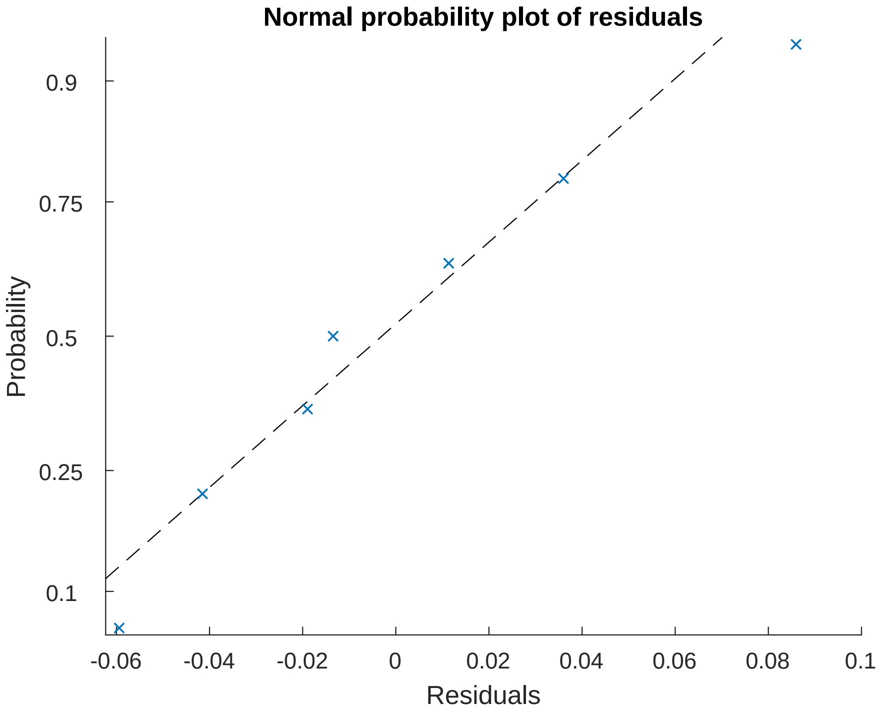
f) Does the simple regression model specify a **useful** relationship between production and protein? Test at level $\alpha = 0.05$ the significance of the predictor X in the model and draw a conclusion.

g) Estimate the true average protein content for all cows whose production is 30 kg/day. Use a confidence interval of 99%.

h) Calculate a 99% **prediction** interval for the protein from a single cow whose production is 30 kg/day.

End of the tutorial class for Week 12. (Graphs on next page)


Linear model output
Extent mist
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
100
200
300
400
500
600
700
800
900
1000
Fluid velocity
Data
Fit
Confidence intervals
Prediction intervals


Normal probability plot of residuals
Probability
0.9
0.75
0.5
0.25
0.1
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08
0.1
Residuals

MATH2099/MATH2859
Probability, Statistics and Information

Statistics Tutorial Class Week 13 : ANOVA

Instructions

The exercises that will be covered during the tutorial class will be chosen from those printed here.

Note that ***not every topic in the course can be covered by the tutorial exercises.*** *You are expected to take an active part in the learning process, by working by yourself on most of the topics.*

In particular, ***you are expected to attempt these tutorial questions beforehand****. At the tutorial, the tutor will go through the answers to some of the questions, directing explanation to areas where students indicate they have difficulty.*

Solutions to these exercises will be available from the course web page later.

Exercise 1

An experiment was carried out to compare the flow rates of 6 different types of nozzles, types A, B, C, D, E and F. ANOVA calculations yielded an observed F value $f_0 = 4.2$ from 66 observations. State H_0 and H_a for the analysis of variance, and carry out the hypothesis test at level $\alpha = 0.05$, giving the range of values for the p-value which can be determined from the tables, and stating your conclusion in plain language. What assumptions need to be made for an ANOVA to be an appropriate test? How could you check if they are plausible assumptions?

Exercise 2*

An experiment was carried out to compare the yield of 4 different crops. Random samples were taken of size 12 from each crop. The observed ms_{Er} was 8.2 and the observed value of the test statistic was $f_0 = 2.8$.

a) Draw up and complete the ANOVA table.

b) Carry out the test using a significance level of $\alpha = 0.01$. Include the statements of H_0 and H_a, rejection criterion, observed value of the test statistic, p-value, and your conclusion in plain language.

Exercise 3*

Having a pet may reduce the owner's stress level. To examine this effect, researchers recruited 45 women who said they were dog lovers. The subjects were randomly assigned to three groups of fifteen women. One group did a stressful task alone; one group with a good friend present; and one group with their dog present. The heart rate during the task is one measure of the effect of stress. Heart rates (beats per minute) during stress were recorded for these 45 women. Below you see the computer ANOVA output for these data.

```
>> stressMod=fitlm(Group,Rate,'CategoricalVars',[1]);
>> anova(stressMod)
ans =

             SumSq     DF    MeanSq      F          pValue
             ------    --    ------    -----      ----------

    x1       2387.7     2    1193.8    14.08      2.0916e-05
    Error    3561.3    42    84.793

>> [meanRate,sdRate]=grpstats(Rate,Group,{'mean','std'});
>> [meanRate,sdRate]
```

```
ans =

   73.4831    9.9698
   91.3251    8.3411
   82.5241    9.2416
```

a) What are the null and alternative hypothesis being tested?
b) From the ANOVA output find the observed value of the F-statistic and the corresponding p-value. State both degrees of freedom for the F-statistic.
c) Briefly describe your conclusions.
d) What are the sample means for the heartrates of the three groups of women in the study?
e) There are three groups considered in this study. If you carry out t-tests to compare them two at a time (pairwise t-tests), how many t-tests are there? Using the Bonferroni adjustment, carry out a t test comparing the group which did a stressful task alone and the group with a good friend present.

Exercise 4

A university development director conducted a survey of starting salaries offered to education majors in three states. Ten offers were recorded for each state, and a partially completed ANOVA table appears below.

Source	df	SS	MS	F
Treatment	**A**	17470016	**B**	**E**
Error	27	**D**	874482.48	
Total	**C**	41081016		

a) Complete the ANOVA table by determining the values for A, B, C, D, E.
b) Test to see if there is a significant difference in mean starting salaries between states at level $\alpha = 0.05$. Clearly state the null and alternative hypothesis both in mathematical terms and in simple language.

That's it for tutorial exercises. You should try to do the rest in MATLAB on your own time.

Exercise 5

Ten batches of concrete were each split into three samples, and a different curing method (1, 2 or 3) was applied to the three samples within each batch. The strength (MPa) was measured for all 30 samples. The data are in the file *concrete.txt*. The aim of the experiment was to investigate differences in mean strength between the different curing methods.

a) Recover the variables `Strength`, `Method`, and `Batch` imported from the *concrete.txt* data set into MATLAB. `Strength` is a measure of the strength of the concrete, `Method` is a categorical variable for the curing method and `Batch` tells you which batch this sample is from.

b) State the null and alternative hypotheses for the appropriate hypothesis test. Which statistical technique do you think is suitable to assess these hypotheses?

c) An ANOVA is a special case of a linear regression, where the `x` variables (predictors) are categorical. We can use the `fitlm` function to carry out an ANOVA on the concrete data. To do this you have to tell the function which variables in your `X` matrix are categorical, using the `'CategoricalVars'` option followed by a vector specifying which columns in `X` are categorical. We then call the `anova` function of the linear model object. For this example try:

```
X=[Method,Batch];
concreteMod=fitlm(X,Strength,'CategoricalVars',[1,2]);
anova(concreteMod)
```

d) Using a 5% significance level, test the hypothesis in b) and state your conclusion.

e) For this ANOVA plot the residuals versus fitted values and a normal quantile plot of residuals and comment.

f) How would your previous conclusion change if `Batch` was omitted from the model?

End of the tutorial class for Week 13.