MAST7866 — Foundations of Data Science

Computing Session 2 — Graphical methods 1

Within this worksheet we are going to look at producing some basic graphs in R.

Task 1 – Graphs

Watch the video "Graphical Methods" on moodle.

Task 2 – Line Charts

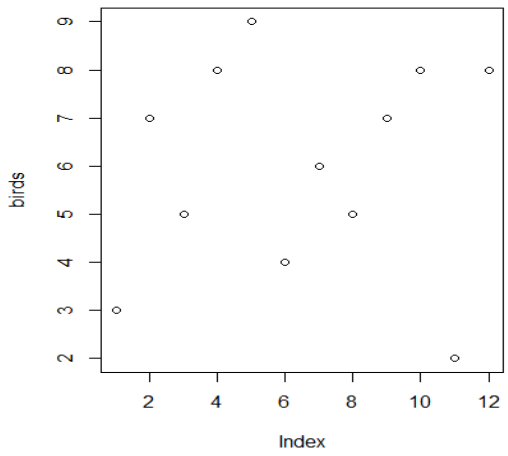
Line charts are useful for displaying time-series data. We will start by producing a very simple graph using the values in a vector called birds. This data set corresponds to the count of the number of birds observed feeding between 8am and 7pm at hourly intervals.

We define the birds vector with 12 values:

```
birds <- c(3, 7, 5, 8, 9, 4, 6, 5, 7, 8, 2, 8)
```

We now produce the graph

```
plot(birds)
```



This graph is not very informative because it has used all of the default function settings. Type `?plot` to find out more about the options within this function.

Suppose we want the x-axis of the graph to denote the time of day the observations were observed. Let us define a new vector:

```
time.of.day <- 8:19
```

Then we can plot time.of.day against birds:

```
plot(time.of.day,birds)
```

We can add a title and a line to connect the points, and some colour as follows:

```
plot(time.of.day,birds, type=”o”, col=”blue”)

title(main=” Number of birds observed feeding”,
col.main=”purple”, font.main=3)
```

If we want to change the x and y-axis labels we have to set these within the plot function:

```
plot(time.of.day,birds, type=”o”, col=”blue”, xlab=”Time of
Day”, ylab=”Number of Birds”)

title(main=” Number of birds observed feeding”,
col.main=”purple”, font.main=3)
```

Now suppose we also have data for cats which we want to show on the same graph.

Let us load the cat data into R:

```
cats <- c(9, 10, 14, 5, 10, 18, 17, 16, 9, 8, 6, 12)
```

Plot the bird line

```
plot(birds, type=”o”, col=”blue”)
```

Use the lines function to add the cat line:

```
lines(cats, type=”o”,pch=22, lty=2, col=”red”)
```

You will see that the cat line is not fully displayed on this plot as the y-axis does not go high enough. This is an annoyance of base R graphics. Adding a 2nd line to a graph you must consider the initial plot to be fixed as an image file and then the 2nd line is added to that image. So you have to pre-set the axes ranges when you plot the initial line:

```
plot(time.of.day,birds, type="o",
col="blue",ylim=c(0,20),ylab="Number of animals",xlab="Time of
Day")

lines(time.of.day,cats, type="o",pch=22, lty=2, col="red")
```

Next we might want to add a legend to let people know what the blue and red lines represent.

```
legend(8, 20, c(“birds”,”cats”), cex=0.8, col=c(“blue”,”red”),
pch=21:22, lty=1:2)
```

The 8 and 20 in this command represent that we want the top left-hand corner of the legend box to sit at the point x=8 and y=20 of the graph.

Challenge 1

Every year researchers count the number of ibex in Gran Paradiso National Park in Italy. The numbers of males and females counted at the park between 1956 and 2016 are provided in ibex_counts.txt which can be found on moodle. Read the data into R and construct a plot showing the data for both males and females.

Note: you might find the function `max` to find out what the limits of the y-axis need to be.

Task 3 – Pie Charts

The file `BirdFluCases.txt` contains the annual number of confirmed cases of human Avian Influenza A (H5N1) for several countries reported to the World Health Organization.

Read the data into R.

We have annual data from the years 2003-2008. The first variable contains the years. There are various things we can learn from this data set. An interesting question is whether the number of bird flu cases has increased over time. We can address this question for individual countries or for the total number of cases. The latter is calculated by

```
Cases <- rowSums(BirdFluCases[,2:16])

names(Cases) <- BirdFluCases[,1]

Cases
```

The name command assigns a label to each data point which will then be used to label the plots you produce. The label we have assigned here is the year the cases correspond to.

Now let us plot some pie charts of this case data

We want to plot 4 graphs on the same display so we use the following command to split the window into 4. Note when you run this command you will not see any change until you plot your graphs.

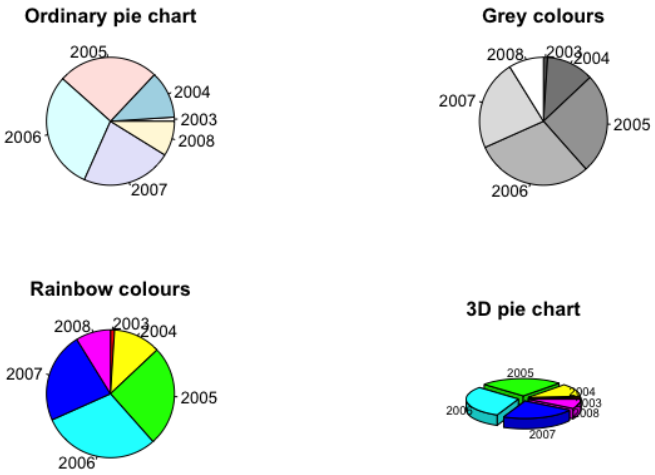
```
par(mfrow = c(2,2), mar = c(3, 3, 2, 1))
```

We are going to plot 4 different pie charts:

```
pie(Cases, main="Ordinary pie chart")

pie(Cases, col=gray(seq(0.4, 1.0, length = 6)),
clockwise=TRUE, main="Grey colours")

pie(Cases, col=rainbow(6), clockwise=TRUE, main="Rainbow
colours")
```

For the last pie chart we are going to use a package called plotrix:

```
install.packages("plotrix")

library(plotrix)

pie3D(Cases, labels=names(Cases),explode=0.1, main="3D pie
chart", labelcex = 0.6)
```



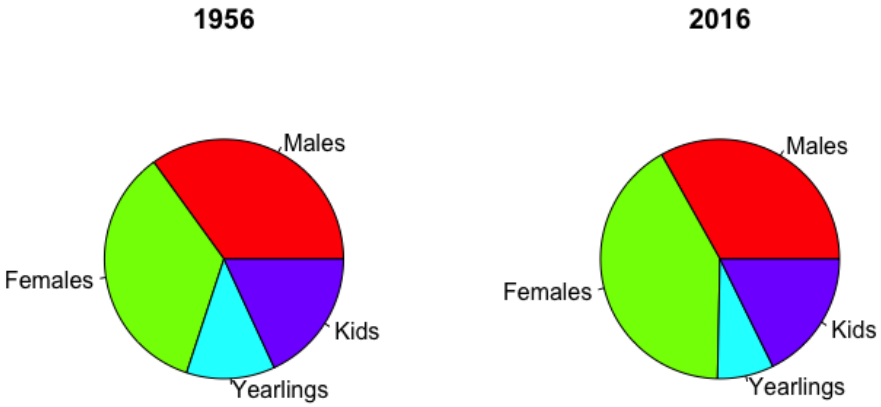
You will notice that although the 3D pie chart is possible, if anything it is less clear in its presentation than the other three graphs. Don't overcomplicate your plots!

Challenge 2

For the ibex population in 1956 there were 681 males, 680 females, 233 yearlings and 352 kids. In 2016 there were 895 males, 1122 females, 206 yearlings and 478 kids. Plot two pie charts side by side to show these data.

You will first need to input the data in the appropriate format:

```
ibex_1956<-c(681,680,233,352)
ibex_2016<-c(895,1122,206,478)
names(ibex_1956)<-c("Males","Females","Yearlings","Kids")
names(ibex_2016)<-c("Males","Females","Yearlings","Kids")
```

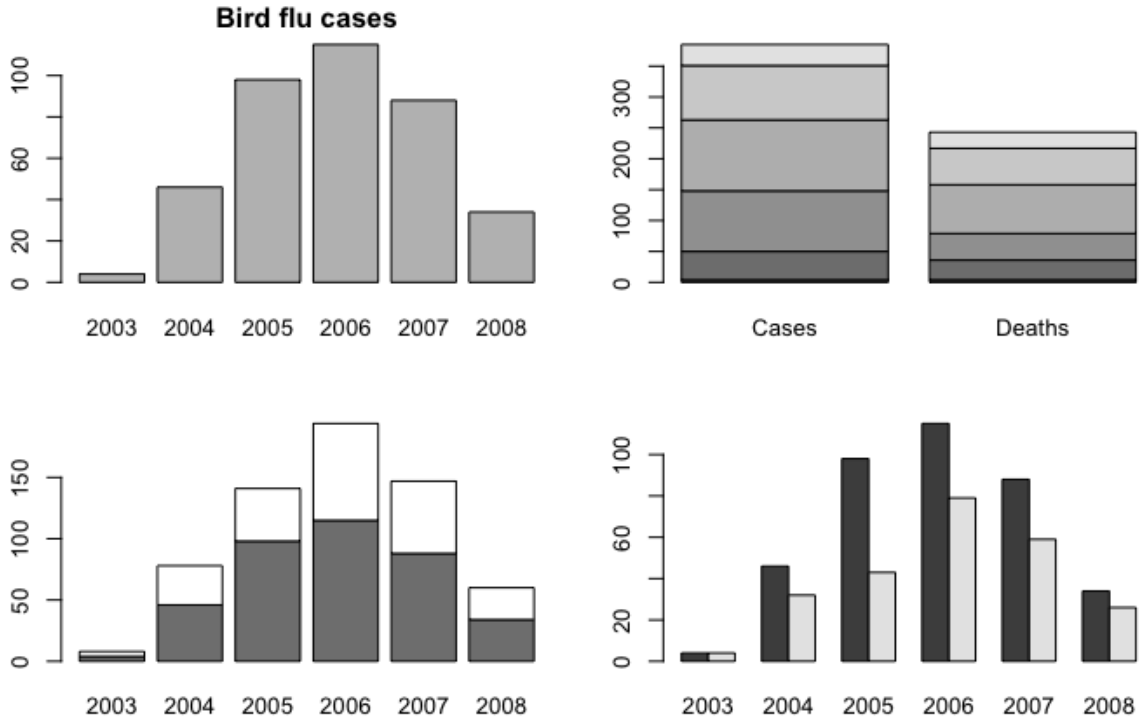


Task 4 – Bar Charts

In the previous section an avian influenza data set was used to create pie charts showing the total number of cases per year. In addition to bird flu cases, the number of deaths is also available (BirdFluDeaths.txt).

```
Deaths <- rowSums(BirdFluDeaths[,2:16])

names(Deaths) <- BirdFluDeaths[,1]

Deaths
```

The data are structured in the same manner as the bird flu cases. We can visualise the change in the number of cases over time, and then compare number of cases to deaths.



These plots are created using the following commands

```
par(mfrow = c(2,2), mar = c(3, 3, 2, 1))

barplot(Cases, main="Bird flu cases")

Counts <- cbind(Cases, Deaths)

barplot(Counts)

barplot(t(Counts), col=gray(c(0.5, 1)))

barplot(t(Counts), beside=TRUE)
```

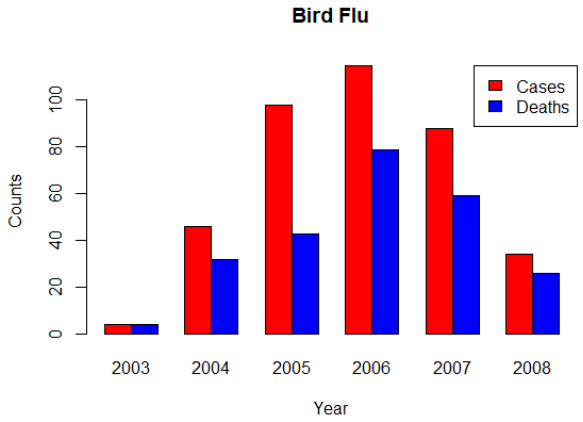
The top left hand plot is a simple bar plot of the number of cases of bird flu over time. In the top right hand plot the bars represent data for each year. The graph gives little useable

information. Also, years with very small numbers, e.g. 2003 are barely visible. In the bottom left hand plot the cases and deaths are stacked and such plots can be misleading. The clearest plot is the one in the bottom right hand plot which shows bars next to each other for each of the years.

Note suppose we just want to plot the final graph and add a legend and axes labels we would use the following commands.

```
par(mfrow = c(1,1),mar = c(5, 5, 5, 5))
barplot(t(Counts), beside=TRUE,col = c("red","blue"))
title(main="Bird Flu",xlab="Year",ylab="Counts")
legend("topright", legend = c("Cases", "Deaths"),fill =
c("red", "blue"))
```

Note that we have had to increase the margin size using the `mar` command in the top line of code in order to be able to see the x and y labels on the axes.



Challenge 3

The file births_maternal_age.txt contains the numbers of live births per age class of women in several years between 1947 and 2017. Plot some appropriate bar charts to compare the numbers of births by age group and by year.