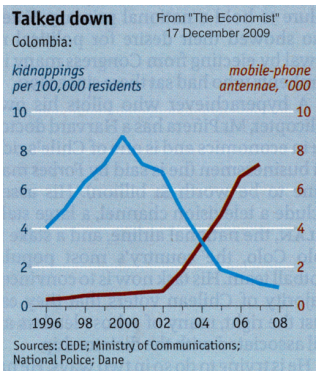
Statistics for Engineers, 4C3/6C3

Assignment 1

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 12 January 2011**

Note: 600-level students must complete all the question; 400-level students may attempt the 600 level question for extra credit.

Question 1 [1.5]



1. What type of plot is this?
2. Describe the phenomenon displayed.
3. Which plot type asks you to draw a cause and effect relationship?
4. Use rough values from the given plot to construct an approximate example of the plot you proposed in part 3.
5. What advantage is there to the plot given here, over the type in your answer to part 3.

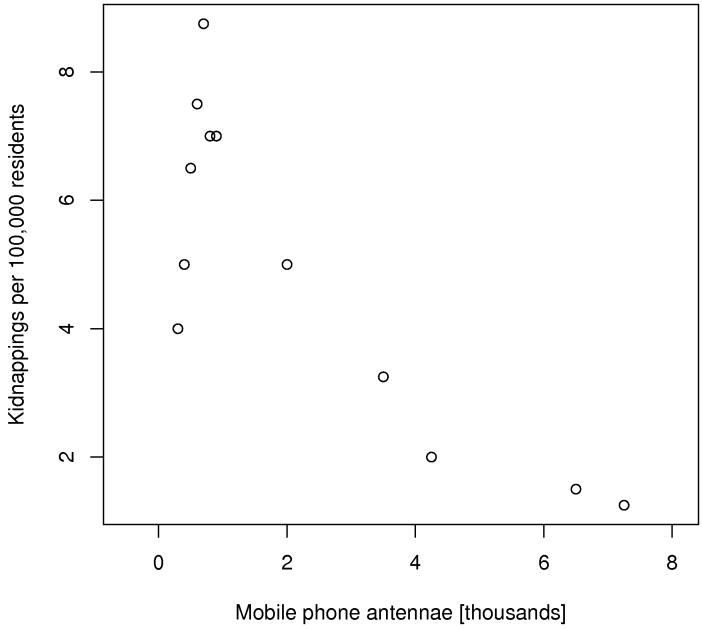
Solution

1. A time-series plot.
2. The rate of cellphone usage (expected to be proportional to number of mobile phone antennae) has increased in Columbia, especially since 2002. Likely this is this usual case where the price comes down, leading to greater use. Though some other political or economic change may have taken place in 2002 leading to increased phone use.

   The rate of kidnappings peaked in 2000, at a rate of 8 per 100,000 residents, and has steadily decreased since that peak.
3. A scatter plot.
4. A scatter plot, from approximate values on the plot, is generated by the following code (you may use any software to construct your plot)

```
# Data from 1996 to 2007
bitmap('../images/kidnap-mobile.jpg', pointsize=14, res=300)
kidnap <- c(  4,   5, 6.5, 7.5, 8.75,   7,   7, 5, 3.25,    2, 1.5, 1.25)
mobile <- c(0.3, 0.4, 0.5, 0.6,  0.7, 0.8, 0.9, 2,  3.5, 4.25, 6.5, 7.25)
```

```
plot(mobile, kidnap, type='p', xlab="Mobile phone antennae [thousands]",
     ylab="Kidnappings per 100,000 residents")
dev.off()
```



5. The advantage of the time-series plot is that you are able to clearly see any time-based trends - those are lost in the scatter plot (though you can recover some time-based information when you connect the dots in time order).

Comment:

The general negative correlation in the scatter plot, and the trends observed in the time-series plots ask you to infer a relationship between the two trajectories. In this case the plot's author would like you to infer that increased cellphone penetration in the population has been (partly) responsible for the reduction in kidnappings.

This relationship may, or may not be, causal in nature. The only way to ascertain causality would be to do an experiment: in this case, you would remove cellphone antennae and see if kidnappings increased again. This example outlines the problem with trends and data observed from society - we can never be sure the phenomena are causal:

- firstly we couldn't possibly perform that experiment of removing cell towers, and
- even if we could, the time scales are too long to control the experimental conditions: something else would change while we were doing the experiment.

To compensate for that, social science studies compare similar countries - for example the original article from The Economist's website shows how the same data from Mexico and Venezuela were compared to Columbia's data. The article also shows how much of the trend was due to political changes in the country that were happening at the same time: in particular a 3rd factor not shown in the plots was largely responsible for the decrease in kidnappings. Kidnappings would probably have remained at the same level if it were not also for the increase in the number of police officers, who are able to respond to citizen's cellphone calls.

Fortunately in engineering situations we deal with much shorter time scales, and are able to better control our experiments. However the case of an uncertain 3rd factor is prevalent and must be guarded for - we'll learn about this is the section on design of experiments.

Question 2 [1.5]

Load the room temperature dataset from the general Datasets website into R, Python or MATLAB.

1. Plot the 4 trajectories, `FrontLeft`, `FrontRight`, `BackLeft` and `BackRight` on the same plot using the default settings in the software.
2. Comment on any features you observe in your plot.
3. Be specific and describe how sparklines of these same data would improve the message the data is showing.

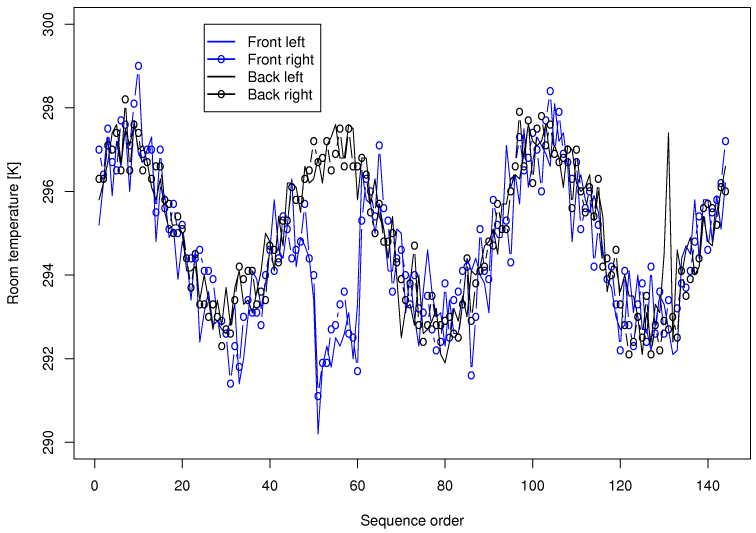
Solution

1. You could use the following code to plot the data:

```
roomtemp <- read.csv('http://datasets.connectmv.com/file/room-temperature.csv')
summary(roomtemp)
ylim = c(290, 300)

bitmap('../images/assgn1-room-temperatures.jpg', pointsize=14, res=300, width=10)
par(mar=c(4, 4, 0.2, 0.2))  # (bottom, left, top, right); defaults are par(mar=c(5, 4, 4, 2) + 0
plot(roomtemp$FrontLeft,   type='l',           col="blue", ylim=ylim, xlab="Sequence order", yla
lines(roomtemp$FrontRight, type='b',  pch='o', col="blue")
lines(roomtemp$BackLeft,   type='l',           col="black")
lines(roomtemp$BackRight,  type='b',  pch='o', col="black")

legend(25, 300, legend=c("Front left", "Front right", "Back left", "Back right"),
       col=c("blue", "blue", "black", "black"), lwd=2, pch=c(NA, "o", NA, "o"))
dev.off()
```



We did not expect you to plot time-based plots: a sequence plot was good enough. There will be a tutorial later showing how to get a time-based x-axis.

2.
   - Oscillations, with a period of roughly 48 to 50 samples (corresponds to 24 hours) shows a daily cycle in the temperature.
   - All 4 temperatures are correlated (move together).
   - There is a break in the correlation around samples 50 to 60 on the front temperatures (maybe a door or window was left open?). Notice that the oscillatory trend still continues within the offset region - just shifted lower.

- A spike up in the room's back left temperature, around sample 135.

3. The above plot was requested to be on one axis, which leads to some clutter in the presentation. Sparklines show each trajectory on their own axis, so it is less cluttered, but the same features would still be observed when the 4 tiny plots are stacked one on top of each other.

   Another example of effective sparklines are for stock market data. Take a look, for example at Google Finance for ERJ (Embraer SA). Google shows Embraer's stock price, but scroll down to see the sparklines for other companies that are in the same economic sector (Bombadier, Boeing, Northrop Grumman, *etc*). This quickly allows you to see whether movements in a stock are due to the overall sector (correlations), or due to a particular company (broken correlations).

   If you looked around for how to generate sparklines in R you may have come across this website. Notice in the top left corner that the `sparklines` function comes from the `YaleToolkit`, which is an add-on package to R. I show how to install packages in the tutorial. Once installed, you can try out that `sparklines` function:

   - First load the library: `library(YaleToolkit)`
   - Then see the help for the function: `help(sparklines)`

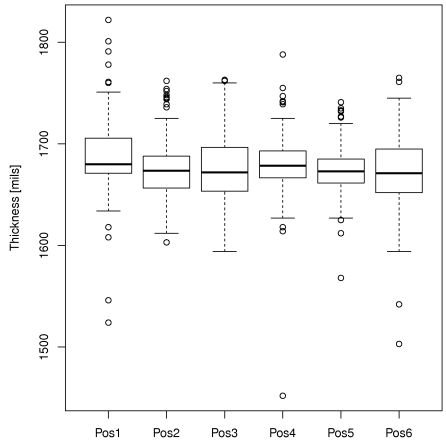
Question 3 [1]

Load the six point board thickness dataset, available from datasets website.

1. Plot a boxplot of the first 100 rows of data to match the figure in the course notes (page 9 in the PDF version).
2. Explain why the thick center line in the box plot is not symmetrical with the outer edges of the box.

Solution

1. The following code will load the data, and plot a boxplot on the first 100 rows:




```
boards <- read.csv('http://datasets.connectmv.com/file/six-point-board-thickness.csv')
summary(boards)

# Ignore the first date/time column: using only Pos1, Pos2, ... Pos6 columns
first100 <- boards[1:100, 2:7]
bitmap('../images/assgn1-thickness-boxplots.jpg', pointsize=14, res=300)
par(mar=c(2, 4, 0.2, 0.2))  # (bottom, left, top, right) spacing around plot
boxplot(first100, ylab="Thickness [mils]")
dev.off()
```

2. The thick center line on each boxplot is the median (50th percentile) of that variable. The top and bottom edges of the box are the 25th and 75th percentile, respectively. If the data are from a symmetric distribution, such as the t or normal distribution, then the median should be approximately centered with respect to those 2 percentiles. The fact that it is not, especially for position 1, indicates the data are *skewed* either to the left (median is closer to upper edge) or the the right (median closer to the lower edge).

Question 4 [1]

Pie charts are widely criticized in the technical literature as being inappropriate - there is almost never a case where it is suitable - yet we see them in the media all the time. Next time you open a daily newspaper or magazine count how many times you see this type of plot.

Read the article by Stephen Few, "Save the pies for dessert" and explain in your own words the shortcomings of the pie chart. Which is an appropriate alternative?

Solution

Please read the article.

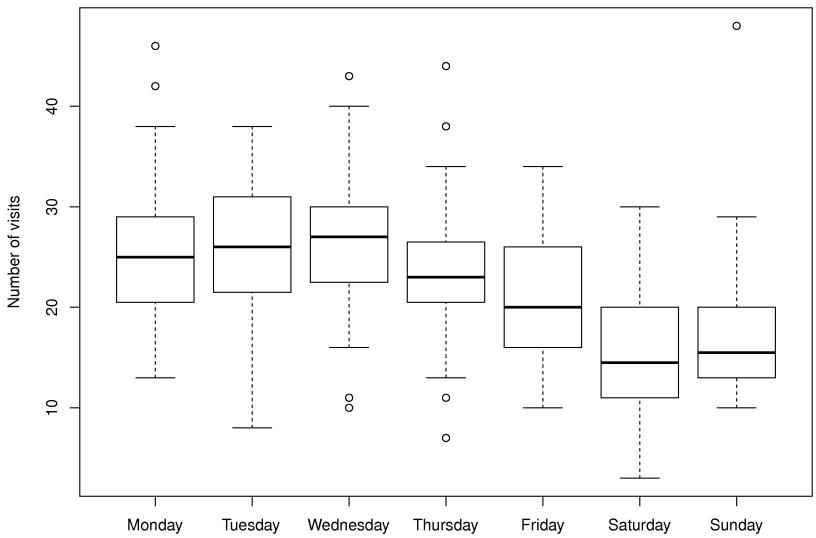
Question 5 [1]

Using the Website traffic data set:

1. Create a chart that shows the *variability* in website traffic for each day of the week.
2. Use the same data set to describe any time-based trends that are apparent.

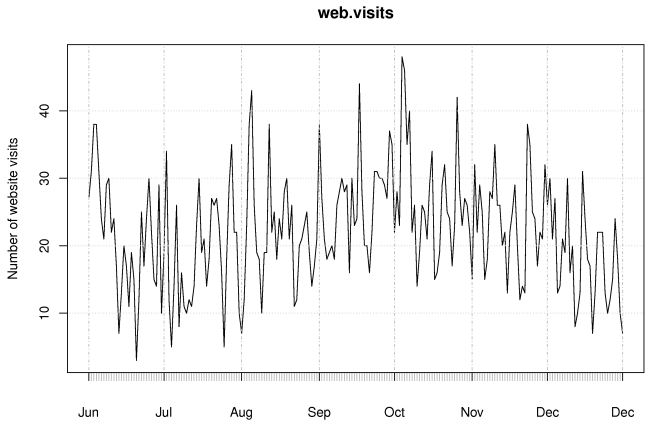
Solution

1. A suitable chart for displaying variability on a per-day basis is the boxplot, one box for each day of the week. This allows you to see *between-day* variation when comparing the boxes side by side, and get an impression of the *variability within* each variable, by examining how the box's horizontal lines are spread out (25th, 50th and 75th percentiles).



We can see much less traffic to the website on weekends. During the week we see a day-to-day increase in the median number of visits, peaking on Wednesday and then dropping off by Friday. All week days seem to have about the same level of spread, except Friday, which is more variable.

2. A time-series plot of these same data is shown here. Apart from the oscillatory trends within the week, we can also observe a general increasing trend in the number of visits during September and October and dropping off again in November and December. If you plot the total visits within each month you can see this effect in a nice way. The lowest number of visits were recorded in late June and July.




```
web <- read.csv('http://datasets.connectmv.com/file/website-traffic.csv')
summary(web)

# Plot the default boxplot: the days are not in the usual order
boxplot(web$Visits ~ web$DayOfWeek)

# Create a factor variable to reorder the days in a new order:
# The factor names MUST match the spelling used in the original file.
day.names <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")
day.factor <- factor(web$DayOfWeek, level=day.names)

# Plot the boxplot in the new order:
bitmap('../images/assgn1-website-traffic-boxplots.jpg', res=300, pointsize=14, width=10)
par(mar=c(4, 4, 0.2, 0.2))  # (bottom, left, top, right) spacing around plot
boxplot(web$Visits ~ day.factor, ylab="Number of visits")
dev.off()

# Plot the data in a time-series plot: crude method - OK for now
plot(web$Visits, type="o")

# A better plot (not expected to get full grade for this question)
# Use the xts library; search the software tutorial for "xts" to see how.
library(xts)
date.order <- as.Date(web$MonthDay, format=" %B %d")
web.visits <- xts(web$Visits, order.by=date.order)
bitmap('../images/assgn1-website-traffic-timeseries.jpg', res=300, pointsize=14, width=10)
plot(web.visits, major.format="%b", ylab="Number of website visits")
dev.off()
```

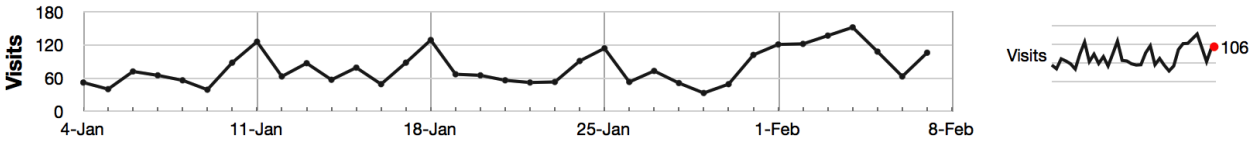
Question 6 [1] (600 level)

Copy a plot from any recently-graduated PhD student in your department. Include the plot in your assignment and comment on any shortcomings in the plot: how would you improve it and if necessary, reproduce your version of the improved plot.

Non-credit question

Note: *Question 1* from the course notes was a question from the 2010 midterm. Repeated below.

The data shown here are the number of visits to a university website for the 4C3/6C3 statistics course in 2010. There were 90 students in the course, however the site is also publicly available.



1. What are the names (type) of the 2 plots shown?
2. List any 2 interesting features in these data.

Solution

1. A time-series plot and a sparkline. The sparkline shows exactly the same data, just a more compact form (without the labelling on the axes).
2. Some of the features shown in the data are:
   - A noticeable weekly cycle; probably assignments are due the day after the peaks.
   - A sustained, high level of traffic in the first week February (a midterm test).
   - Some days have more than 90 visits, indicating that students visit the site more than once per day, or due to external visitors to the site.

END