1. Introduction

Statistics is taught in many academic disciplines. It can be taught in a theoretical way (as with courses in mathematical statistics) and it can be taught in an applied way (as happens in most other courses). Quite what is meant by applied varies from course to course, not only in the area of applications (psychological examples may not be as attractive to geography students as they are to psychology students), but also in what is meant by applied. When the author of this report first learnt statistics, examples often amounted to little more than calculations with a few numbers. Sometimes the numbers were even embellished with fictional background (“Factory A produces...”). Real examples became more common over the years, though they were often overused - the iris dataset being a prime example. Thanks to the internet, there are now many more opportunities for finding interesting, up-to-date, understandable examples with real data. The purpose of the workshop was to discuss this new resource in terms of some basic questions:

1. Is it a good idea to use real datasets in statistics teaching?
2. What are the advantages and disadvantages?
3. What web resources are there?

Additionally, we were interested in how applications involving statistics are discussed on the web. There are plenty of examples of discussions of data, which would benefit from a statistical contribution. Should statisticians get more involved? Or should we just be pleased that discussions of data are taking place at all and hope that this open-ended discussion will lead to general, if slow, progress?

2. Using real datasets in teaching statistics

A workshop organised to discuss teaching with datasets is not going to attract participants who do not believe in teaching with datasets. The tenor of the meeting included an implicit agreement that real datasets were a good thing, combined with the common experience that it was difficult to find ideal examples, and the general agreement that teaching of this kind could be improved. Clearly the possibilities offered by the web should be investigated in more depth.

2.1 Advantages

Real datasets are motivating, students can see how what they are learning can be applied. Statistics can be useful in just about every field and so it is possible to find good examples for all kinds of interests. Using a wide range of attractive examples
may also make students realise that statistics is actually a fascinating area to work in and not the dry and boring subject that it is viewed as by some. Real datasets illustrate issues that arise in practice, such as data quality, outliers, and missing values. They show the importance and relevance of assumptions. Requiring that a variable be normally distributed as part of the proof of a theorem is very different from thinking about whether some data is normally distributed in practice and evaluating how important or not that might be. This is part of putting theory into context, not presenting it only as an intellectual structure, but also as part of a methodology that can be applied in practice. Working with real examples helps the students gain practical experience, which makes them better able to face the world they have to deal with after university. Statistics is a part of so much of everyday life nowadays, whether we are referring to opinion polls, the results of medical studies, demographic information, or whatever, that an education in Statistics which restricts itself to theory can only be regarded as incomplete.

2.2 Disadvantages

Using a real example in class generally requires a lot of preparation. You need to understand the aims of the study and how it was carried out. You often need to reformat and tidy up the data to make it suitable for use with students. Studies you have worked on yourself in a consultancy role with other researchers are good in this respect, because you already know all the background. They may not be so good if the application takes a long time to explain and if the topic is not of immediate interest. Real studies often have special features, which may distract from the statistical issues you want to illustrate, and sometimes the very topic may awaken students’ interest to such an extent, that statistical concepts and theory get left by the wayside. And, of course, it is difficult to find examples that interest everyone. The topics of some examples may simply leave many students cold: why should everyone be interested in sports or politics or physics? The fact that preparing real examples for teaching takes up a lot of time raises another problem, is there enough time on the course? It has been suggested that students can gather practical experience at university and there is not enough time to give them a proper statistical education as it is.

3. Sources of datasets

In the past textbooks were one of the few sources of datasets. There were two books devoted only to datasets. Andrews and Herzberg included mainly historical examples and was notable for the authors’ efforts to explain the background to the data. Hand et al.’s collection gave less information, but included many more datasets. Some textbooks use real datasets and often make them available on websites. A particularly attractive collection of real examples for teaching is associated with Ramsey and Schafer’s text “Statistical Sleuth” (http://www.proxaxis.com/~panorama/home.htm).

Nowadays we look primarily on the web. There are sources of different kinds and one possible classification is:

1. Government and public bodies;
2. Media (both news media and academic journals);
3. Commercial and other sites;
4. Academic;
5. Data visualisation sites; and,
6. Blogs

The following paragraphs give some details of typical sites of each kind. The lists are incomplete and make no claims to have found the best current sites. Hopefully, they give useful impressions of the kinds of sites that exist.

3.1 Government and public bodies

Both the UK (http://data.gov.uk) and US (http://www.data.gov) governments have developed websites making statistics more accessible ways in recent years. The Statistisches Bundesamt in Germany (http://www.destatis.de) also has a substantial website. The NHS in Scotland have the site http://www.isdscotland.org and there are many other public organisations providing statistical data. Mostly they report results rather than provide data. Many tables may be available, but not the multivariate data underlying them.

3.2 Media (both news media and academic journals)

The Guardian newspaper website in the UK includes a data store (http://www.guardian.co.uk/data-store), which offers a range of datasets of public interest, carefully prepared for use. The choice of datasets is partly determined by current events and partly, doubtless, by availability. Some academic journals now offer authors the opportunity to provide background materials for their papers, including datasets, on the web. The Royal Statistical Society (RSS) (http://www.blackwellpublishing.com/rss/) does this for its journals, though up until now this facility has not been used extensively. The American Statistical Association (ASA) has a clear policy on authors’ making data available: “Whenever a dataset is used, its source should be fully documented. When it is not practical to include the whole of a dataset in the paper, the paper should state how the complete dataset can be obtained. Unless conditions of security or confidentiality intervene, availability of the data on which the paper is based is a requirement for publication.” (http://pubs.amstat.org/page/jasa/information-for-authors) There is a repository at StatLib (http://lib.stat.cmu.edu/jasadata/).

3.3 Commercial and other sites

Both Amazon (http://aws.amazon.com/publicdatasets/) and Google (http://www.google.com/publicdata/) now offer repositories of public data. At the time of writing
these are still in their infancy and it will be interesting to see how or if they develop. In Germany there is a site Statista (http://de.statista.com) which claims to combine statistics from over 500 different sources. Some of their material is free; some can only be obtained by paying subscribers. Only aggregated data is available. The film website IMDb (http://www.imdb.com) allows users to rate films they have seen. Data by film, broken down by population groups, is available by individual film. For sporting data there are many sites (e.g. http://www.decathlon2000.ee or http://basketballvalue.com). Some are commercial; some are run as a service. For political roll-call data there is a site at Berkeley (http://voteworld.berkeley.edu). For climate data there are many sites, for instance: http://iridl.ldeo.columbia.edu or http://www.cru.uea.ac.uk/cru/data. There are many other data sources for other topics on the web. Much depends on the enthusiasm and resources of the site maintainers.

3.4 Academic

DASL (Data And Story Library, http://lib.stat.cmu.edu/DASL/) was an innovative site started in the mid 1990's at Carnegie-Mellon in Pittsburgh. It offered small datasets with accompanying explanations and was cross-referenced by application area and statistical methods. Although still online it has not been updated for a while. Statsci.org, set up and maintained by Gordon Smyth in Melbourne, includes many links to dataset sources and is a useful starting point. The Maths, Stats and OR Network also offers some statistics resources and links to datasets (http://www.mathstore.ac.uk/?q=node/1455). Other individual academics or departments offer datasets or collections of links to material. Some of these sites are more permanent than others.

The machine learning community has a much used repository of data sets at UCI Irvine (http://archive.ics.uci.edu/ml/). The aim is “for the empirical analysis of machine learning algorithms” rather than to provide datasets for analysis.

R and its packages (http://www.R-project.org/) contain many datasets. Since almost all these have been contributed by academics and researchers, it would be nice to be able to report that they are provided with full associated information and are analysed instructively in the examples. In practice the range of quality of dataset presentation is as broad as the range of quality of the packages. Some are excellent, some are not. Caveat downloader!

3.5 Data visualisation sites

A number of websites have sprung up, which concentrate on visualisation, especially data visualisation. These include Many Eyes (http://manyeyes.alphaworks.ibm.com/manyeyes), which was started by Marty Wattenberg, who designed the entertaining and excellent Namevoyager (http://www.babynamewizard.com/voyager). His idea with Many Eyes was to encourage people to present their own visualisations and to upload their datasets to permit others to present alternatives: “Our goal is to ‘democratize’ visualization and to enable a new social kind of data analysis”. The standard of visualisations on the site and the quality of the datasets uploaded has been very variable.

Some other data visualisation sites like Junk Charts (http://junkcharts.typepad.com) and Flowing Data (http://flowingdata.com) are blogs where interesting discussions take place and sometimes the data involved is made available.

3.6 Blogs

There are several blogs which discuss topics in which data analysis plays a key part, and they usually give sources, though the datasets themselves are not often available. Examples include Freakonomics (http://freakonomics.blogs.nytimes.com), Statistical Modeling, Causal Inference, and Social Science (http://www.stat.columbia.edu/~gelman/blog), and Bad Science (http://www.badscience.net).

4. Open data

Last year Tim Berners-Lee made an impassioned presentation at the TED conference asking for people to put more “raw data” on the web, just as in the past they put documents up on the web. This year, at the corresponding TED meeting in February, he talked about what has happened in the last year (a six minute video of his talk can be viewed on YouTube). Berners-Lee is very pleased with the response to his request and showed some fascinating examples. Possibly because examples with maps are readily convincing and easy to show, all his examples involved geographic mashups. They also had little in the way of statistical input and this especially affected his first example, a map of cycling accidents in the UK that appeared on the Times Labs Blog, two days after his TED talk, on 11th March 2009. As several commenters on the blog pointed out, numbers of accidents are dependent on traffic, accidents vary in severity, and not all accidents are reported. The first criticism is the crucial one. It is disappointing that Berners-Lee ignored it in his talk this year and disappointing, though not surprising, that no one has added the necessary information on cycling traffic, as it would be difficult to get good data.

In general, it is much better that people are encouraged to make data available than that they are criticised for not interpreting it properly. In particular, some of the discussions of data you find on the web can set your teeth on edge.

5. Getting data and possible problems with data

Just because data is available does not mean that you can easily access it in a format you like. Some data may be
provided in one or more of a number of standard formats that can be read by most people (csv, tab-separated, spreadsheet...). Some is presented in pdf format, which needs additional work to convert it to a usable form. In R packages datasets are occasionally provided as tabular results, though they can always be transformed within R. This is a general issue to be dealt with, even if data is readable you may prefer it organised in a different structural form. Then there is the usual checking for text errors, special codes (e.g., for missing values), data errors and other data preparation tasks. The inclination of some software packages to show how clever they are by automatically recoding certain data types (e.g., dates) probably does not need to be mentioned here.

Sometimes data is available only in small chunks, one at a time and not as complete datasets. Writing scripts, for instance in Perl, to automatically extract the data may or may not be viewed favourably by the site in question.

Once the data is downloaded, there remains the question of whether enough background information is available on the reasons why the data was originally collected (which often affects what data was collected and how), how the variables were defined (e.g., income, support for a political party, petrol consumption), how the data was collected, and, a point often neglected, what cleaning, filtering and selection has already taken place.

6. Discussion of data on the web

Surely no statistician disputes that interpreting statistics is a difficult undertaking. Whether others share that view is debatable. The greater availability of information of all kinds on the web encourages people to look for information themselves. The ease with which comments can be contributed on the web leads to far greater publication of comments than has been true previously. The interpretation of data and statistics on the web is often pretty bad, but that is merely a reflection of the general use of data and statistics. Whether it is really worse than what we find elsewhere would be hard to determine. You could argue that the web has made data available to everyone in the way that statistical software made statistical methods available to all researchers in the past. Just as then there was no longer a need to consult a statistician, just dump the data into your software, now you can play around with data and draw your own conclusions without any statistical input. Pessimistically this will lead to an even worse situation (in a variant of Gresham's Law: “Bad analyses drive out good analyses”). Optimistically this will lead to an appreciation of the importance of statistical understanding. Either way it suggests a greater need for statistical education.

7. Final comments

Availability of data alone is not enough, there has to be sufficient information about a dataset to make it usable by others. It would be valuable to establish consensus guidelines to aid people making datasets available. Just what information and how much will vary from dataset to dataset, though a bare minimum would seem to be a reference to the original source, the aims, a description of how the data was collected, and definitions of the variables.

Using real examples in teaching is hard work. You need to choose the examples carefully. They should be interesting and up-to-date. They should be varied and yet you should know a lot about each one. And then you have to prepare the data so that they can be readily used by the students. Is it any wonder that real examples of this kind are not used a great deal?