STATISTICS IN A DAY

Rodney Carr, School of Management Information Systems,
Deakin University, Australia

We explain how people have been successfully taught to carry out good, solid basic statistical analysis of data in a day (7 hours). We describe a workshop called "Statistics in a Day" that has formed a component in regular courses up to "research methods" courses for PhD students and has been run many times as a professional development offering to people from various organizations. In "Statistics in a Day" we teach the Data Analysis Algorithm directly (avoiding much of the underlying theory and terminology) and use MS Excel with XLStatistics to quickly and efficiently illustrate key ideas (as well as for the number-crunching, etc).

How can people master the techniques for statistical analysis of data *in a day*? Before answering this question, let me point out that the title "Statistics in a Day" is (deliberately) misleading it would be better called "Data Analysis in a Day" (but that's not as catchy!). And the answer to the question is that there is a "data analysis algorithm" - in the "Statistics in a Day" workshop we teach this algorithm. The algorithm is quite well-known - it is given in back of many statistics texts where it appears as a flow chart or selection guide. However, it usually is not possible to teach the algorithm directly because, as it is usually presented, it is appears too complicated and contains a considerable amount of fairly technical language. Most computer packages for statistical analysis do not help much, either - the same problem - they (usually) are too complicated and use a lot of technical terminology. And, to make matters worse, most courses on statistics contain more than is actually needed if all that is required is for people to be able to carry out analysis of data (typically, for example, special probability distributions are covered, or details on the Central Limit Theorem are given).

TO TEACH STATISTICS (DATA ANALYSIS) IN A DAY QUICKLY:

- Most of the terminology, technical details and formulas do not need to be known in order to carry out a statistical analysis of data - leave them out (people do, though, need to know some terminology and enough of the underlying theory to appreciate underlying assumptions for tests, etc)

- Leave out much of the underlying probability - whatever is not needed

- Use a simple computing tool. This is essential - you need a tool that can carry out all the necessary computations *without all the terminology getting in the way* and that can be used to *illustrate the concepts quickly*. We use Microsoft Excel with XLStatistics - XLStatistics is a set of workbooks for statistical analysis of data that effectively implements the data analysis algorithm.

SOME SIMPLE STEPS FOR DATA ANALYSIS.

In "Statistics in a Day" we spend the first 3 or so hours going through a series of 6 steps that implement the data analysis algorithm. We use Excel and XLStatistics to carry out most of the number-crunching - some of the steps (at least as they are stated below) are actually fairly specific to Excel and XLStatistics. In the 3 hours, the steps are presented and associated ideas and some essential terminology and ideas discussed "on the fly" via a couple of examples. Here are the steps - these are discussed separately below:

Step 0. Formulate the problem clearly.

Step 1. Define variables. Identify the number of them and their type. Organize the data.

Step 2. Select and open the relevant XLStatistics workbook and copy or link the data into the workbook's Data area.

Step 3. Examine the data descriptively.

Step 4. (Re-)define the problem in terms of appropriate population parameters.

Step 5. Select the appropriate results from the Tests sheet in the XLStatistics workbook. Check assumptions.

Step 6. Interpretation. Write report.

Step 0. Formulate the problem clearly.

This is not actually part of the data analysis algorithm, but has to be discussed because it is impossible to carry out analysis without knowing what you want. Here we also talk briefly about a couple of other important issues including

- Identifying sources of variation versus Reducing the number of variables to simply the problem
- Quality of data (≈ "you can only generalize to the population from which the sample is drawn")

Step 1. Define variables. Identify the number of them and their type. Organize the data.

The data analysis algorithm is basically driven by the number and type of variables involved in the problem at hand. Step 1 “forces” the issue by getting people to clearly identify their variables. There are two basic types of variables that people need to be made aware of: Continuous (quantitative) and Categorical (qualitative). In this step we also talk about the appropriate way that data needs to be organized in a spreadsheet for statistical analysis (variables ↔ columns, cases ↔ rows).

Step 2. Select and open the relevant XLStatistics workbook and copy or link the data into the workbook’s Data area

Data analysis is largely driven by the number and type of variables present (which were identified in Step 1). And XLStatistics is designed with separate workbooks that each contain the analyses appropriate for a given combination of variables (1 continuous, 1 categorical, 1 continuous and 1 categorical, 2 continuous, etc); in this step users simply open the relevant workbook and paste or link their data in (the workbooks each have a special Data area - and there are simple ways of copying or linking data). There is really not much to do in this step, but it is a critical step if using XLStatistics - the data must be pasted into the correct workbook. In a sense, this step moves the common “What analysis do I use” question back to “Identify the variables”, which seems to be something even people with little experience can easily handle. (The “What analysis do I use” question now vanishes because the appropriate analyses are automatically carried out in the XLStatistics workbooks - choices are few after the variables have been identified.)

Step 3. Examine the data descriptively

This is a very important step, but is often overlooked by people carrying out analysis of data - in the XLStatistics workbooks the relevant summaries are produced automatically so people have really to consciously ignore them to do the wrong thing! The basic idea in this step is that “strange” patterns in the data (such as outliers, skewness, high leverage points, etc, etc) need to be detected and appropriate actions taken (for example, if data is badly skewed the mean might not be an appropriate measure of average - maybe the data should be transformed or the median used). In “Statistics in a Day” we play many “what if” games using the XLStatistics workbooks (together with a

couple of special demonstration workbooks) and illustrate (quickly and without the formulas getting in the way) what the various diagrams and measures show and how they can be used.

Step 4. (Re-)define the problem in terms of appropriate population parameters

It is, of course, usually not the sample results that are of interest (they vary from sample-to-sample) - we normally want to generalize to the entire population. The precise formulation of this is done in this step (statisticians often recommended that this actually be done in Step 0 (formulation) - fine, but in practice this can be difficult - if not impossible - without first looking at the data (Step 3)). In “Statistics in a Day” we spend a short time on this step making the point that sample statistics *do* vary from sample to sample via a couple of quite convincing demonstrations. We do not go into details of, say, the distribution of sample means (because precise details are not actually needed for the subsequent analysis), but we do make it clear that they vary and that the estimates become tighter as sample size increases. In the next step we handle the problem of things happening by chance...

Step 5. Select the appropriate results from the Tests sheet in the XLStatistics workbook

After Step 4, people appreciate that some more analysis needs to be carried out in order to handle the problem that things (in a sample) can happen by chance alone. In Step 5 we show how this can be done. The precise tests, etc, that need to be done depend (as previously) to a large extent on the number and type of variables present - the XLStatistics workbooks already contain the appropriate results (on specially-provided worksheets). People just need to be taught the (small) amount of associated terminology so they can select (and modify if desired) the analysis they require. This step covers, again via a number of “what-if” games and demonstrations, the key ideas for:

- Hypothesis tests. We discuss the important ideas and associated terminology (including a good discussion of the power of tests and sample-size determination). P-values are used (they are easy to interpret and relate directly to the problem of “things happening by chance alone”).
- Confidence intervals. We discuss one-and two-sided intervals, the relationship with hypothesis tests and sample-size determination.

Quite advanced tests are looked at (for example, Analysis of Covariance with interaction terms, non-parametric tests, etc) - with XLStatistics such things are quite easy to do because the relevant tests are in the workbooks already without complicated terminology and with all of the technical details and formulas “hidden” away in Excel formulas. We have a good discussion of the underlying assumptions of the tests (using residuals plots, for example) and examine the sensitivity of results resulting from small changes in the data.

Step 6. Interpretation. Write report

This step takes about 15 seconds (!) in “Statistics in a Day” - after Step 5 it is just a matter of selecting whatever is appropriate for incorporation into a report.

THE AFTERNOON’S EXERCISES.

The general discussion of Steps 1-6 outlined above takes up the morning of “Statistics in a Day”. The afternoon is spent in a workshop where people are guided through analyses in a variety of different situations. Steps 1-5 are followed, but, to save time, the report-writing (Step 6) is usually done verbally. Some Excel-type details are discussed (like, for example, the “problem” of pasting charts from Excel into Word as pictures and not chart objects). We also cover a couple of concepts not explicitly covered in the morning’s work (like the difference between random and fixed effects, or backwards regression, etc). The exercises are usually tailored to the particular audience (this is actually quite easy to do - the notes for “Statistics in Day” have been designed so that examples can be replaced by others without affecting the bulk of the notes).

CONCLUSION

It is quite difficult to describe the details of many of the methods used in “Statistics in a Day” because much of the work is visual and interactive. But hopefully the above description gives at least some idea of this successful and very popular course - it really does seem to give people with little or no statistical training the basic important ideas of data analysis - enough for them to be able to carry out good solid analysis of data on their own.