Data Analysis Workshop as a Statistics Teaching Aid

T Krishnan - Calcutta, India

1. Introduction

Most academic departments of statistics in universities and research organisations, besides their courses for training professional statisticians, conduct summer schools and other short user-oriented courses for various types of scientists, to acquaint them with data analysis techniques. We in the Indian Statistical Institute have also been doing so for over three decades. We have conducted such user-oriented courses on statistics at elementary levels covering the basics as well as at somewhat more advanced levels on specialised topics such as Multivariate Analysis, Design and Analysis of Experiments, Time Series Analysis and Forecasting, Survey Techniques, etc. There have been variants of such courses in the form of computer-oriented courses with the use of micro-computers as well as mainframe computers, with program packages. The classes of users have been Biologists, Medical Scientists, Economists, Social Scientists, Psychologists, Technologists from industry, etc. These courses have been along fairly conventional lines with lectures, practical sessions and hands-on computer sessions explaining the concepts and methodology of statistics and data analysis, with illustrations and exercises. These courses have been useful though with limitations of the type discussed below.

Courses of this kind are methodology-oriented and live problems are rarely taken up and formulated. Also, examples given are limited to already well formulated problems, and problems where the particular statistical techniques have been applied successfully. Further, these examples invariably are from a wide variety of fields with most of which a given participant is not always familiar. Thus the most important and difficult aspects of data analysis, namely problem formulation, choice of methods, etc. are rarely discussed in such programmes. Hence the participants do not get an appreciation of the difficulties involved in tackling a quantitative problem and the "trials and errors" involved in analysing data. Invariably the participants in such programmes

find the courses too abstract and as a consequence do not get a good appreciation of the nature of data analysis. On the other hand, statisticians, especially those working in organisations like the Indian Statistical Institute, have been involved in consultation and analysis of data relating to scientific problems from diverse fields. In most of these consultancy sessions, the scientists concerned participate in a limited way; after an initial discussion of scientific problems, the data etc., the statistician works on his/her own and "delivers" the "output" with some remarks, and the scientist perhaps includes these in his/her research papers. Indeed, there are exceptions to this pattern; there have been collaboration studies with full interaction between the scientist and the statistician at all stages of the study, but these are not generally carried out as a programme in which other scientists participate and benefit. Thus typically the consultancy sessions are not used for any kind of training of the scientists, nor does it happen that other statisticians get to know about this work even if it is a novel application of statistical methods, except perhaps in cases where the work results in the development of new and publishable statistical methodology.

A Data Analysis Workshop proposed and discussed here may be a useful device to partly offset these disadvantages of conventional training progammes and consultancy sessions and to combine some of their advantages. Here the idea is to take up a few live data analysis problems from the same area of science, go through the processes of formulating the scientific problems statistically, analysing the data and interpreting the results - all in the workshop itself.

Data Analysis Workshops have been and are being conducted elsewhere. The ones I have seen, attended, and am aware of, seem to be of a different kind from what I am proposing here. For instance, there are Data Analysis Workshops where the participants are entirely statisticians discussing mainly methodological problems. There are Data Analysis Workshops organised on the basis of specific data sets - in these, data sets with write-ups on them are circulated a few months ahead of the programme; participants, who are mainly statisticians, present their results of the analysis with methodological details in the workshop lasting three or four days. In such workshops no analysis is actually carried out during the workshop, nor are scientists whose data sets are analysed involved in the workshop in any significant manner. In the workshop I am proposing here, scientists and statisticians are both equally involved, problems are explained and formulated, analysis explained and carried out and interpretation made - all as a part of the workshop.

I give below the details of the structure of such a workshop using the example of one such workshop conducted at the Indian Statistical Institute.

2. Organisation of the workshop

The first organisational task of such a workshop would be to choose and fix a subject matter and a group of scientists working on quantitative aspects of this subject with sufficiently interesting data sets, which are feasible to be analysed in a period of three weeks. In our workshop, we began with an all-India advertisement announcing a Data Analysis Workshop for the Life Sciences; intending participants were asked to send a brief write-up on an investigation for which they would like data analysis to be undertaken in the workshop. They were asked to include in the write-up the aims and

objectives of the investigation, its design, examples of data, examples of important questions the investigation sought to answer, etc. A proforma for project description was asked to be filled in. On the basis of the write-ups and information provided in the proforma, attempts were made to select six to eight participants. We selected mainly Marine Biologists; however, due to late withdrawal of some selected candidates, we had to find substitutes who were from somewhat different fields, albeit within the broad area of the Life Sciences.

Having selected the participants, the next task was to get the data sets with adequate descriptions so that transfer of data to a medium suitable for use on the computer, scrutiny, and some preliminary studies on them could be done before the commencement of the workshop. In our workshop we asked the selected participants to send their complete data sets two months before the workshop was due to begin. On receipt of these data sets, for each participant we set up a team consisting of about five persons, which included statisticians, computer programmers and a specialist on the subject matter of the investigation, one of the team members acting as coordinator. The respective teams started preparing computer data files and entered into correspondence with the participant asking him/her to sort out difficulties; they also started studying the data sets; some teams carried out some preliminary data analysis as well. The participants were also asked to send in advance, notes for a few lectures that they were expected to deliver in the workshop on the background scientific material required for a reasonable understanding of their scientific investigations. Relevant published reference material was also sought. Handouts were made ready on the basis of this material, to be distributed when the workshop began. On the basis of the preliminary study of the data sets, a list of statistical topics for lectures and tutorial sessions was also worked out. The data sets were sufficiently varied in respect of data analysis requirements, and the list of topics made up a fairly complete course on data analysis at an elementary level.

3. Structure of the workshop

The main features of the workshop were:

(i) presentation of their investigations by the participants;
(ii) lectures by statisticians on methodology of common interest;
(iii) discussion sessions for suggesting approaches to data collection and analysis, reporting progress and interpretation of results;
(iv) consultancy sessions between individual participants and their teams;
(v) data analysis sessions by the scientists.

The first week of the workshop consisted mainly of introduction of the problem areas with background scientific material, explanation of the aims and objectives of the investigations, description of data sets, etc. by the participants. Each such introduction was followed by initial remarks on the data set and tentative protocols for analysis by the statistical teams concerned. On the basis of these sessions and preliminary studies on the data sets, the list of topics prepared earlier was revised. Material for lectures and tutorial sessions was prepared based on the problems and data sets of the workshop. The topics on which these lectures and tutorial sessions were organised are: (1) tests of

significance; (2) regression and correlation analysis; (3) design of experiments; (4) analysis of variance; (5) multivariate analysis; (6) statistical classification. There were also general lectures on statistical methodology and the use of computers including the use of standard statistical packages.

The second week consisted mainly of actual data analysis work in which the participants took part as much as possible. Details of the data analysis were discussed by the team with the participant concerned, methods were explained to him/her, the participant was asked to carry out as much of the work himself/herself as possible and all work on his/her data set was done with him/her. On each day there was a session in which progress was reported, comments made on the analysis performed thus far and suggestions made for alternative approaches etc. During this period lectures and tutorial sessions were also held.

During the third week of the programme results of analysis were presented by each team of statisticians followed by remarks by the participants concerned giving their interpretations in scientific terms for the conclusions reached, expressing satisfaction or otherwise with the results obtained, and asking fresh questions on the basis of the results, etc.

4. Concluding remarks

Although the structure of the workshop as outlined above may appear neat and interesting, it is hardly easy to implement it. A practical problem presents its own innumerable difficulties and in our workshop we had plenty of these. I shall enumerate a few of them.

(i) Although we had originally selected six data sets on very similar topics, in view of the late withdrawal of some participants, we had to substitute them by participants from somewhat different fields. This certainly weakened the workshop. This was clearly evident from the deliberations of the workshop. The most lively discussions - both subject matter and statistical - were certainly on those topics which were common to many participants.

(ii) Despite our best efforts, not all participants provided enough relevant information about their investigations, problems and data sets before the workshop. This made it difficult to make an ideal selection of participants.

(iii) Not all data sets arrived in good time for the data analysis teams to have had a good look at the data before the workshop began.

(iv) Even with the data sets that arrived in time, there were difficulties with their interpretations, since the participants did not know what sort of information was relevant; these concerned missing observations, the meaning of "zero" observations, over-summarisation of data, for example, sending means instead of the raw data, etc.

(v) There were differences in the progress of the various teams, partly owing to the nature of the analysis they were carrying out and partly owing to their own abilities. This resulted in a certain asymmetry in the programme and its rescheduling, especially the "Progress Reporting" part.

Most of these difficulties could be surmounted by more careful planning, which we did in our subsequent workshops. However, despite these difficulties, the workshop was quite successful. The most lively and interesting part was the first part where the participants presented their problems and data sets. There was all-round participation from both the statisticians and the scientists in the discussions. The part consisting of lectures on various topics, in which the various participants' data sets were used to explain statistical concepts and methodology, was also quite successful. However, the progress reporting sessions had their own difficulties owing to reasons explained earlier.

One of the fallouts of the workshop was that the statisticians and the scientists were able to see a quantitative problem from each others' point of view and the participants were able to appreciate the fact that data analysis was not a straight-jacket affair and that it involved looking at the data from various points of view and that even with a very efficient computer facility it does take quite some time to get a data set analysed. Further, the participants learnt the importance of planning their data collection more carefully and learnt the basic principles of experimental designs. They also learnt how to organise their data for analysis on a computer (aspects such as coding, missing values, differentiation between zero, missing value, etc.). They appreciated the importance of representing their data and results by means of plots, graphs and charts. They learnt the importance of preliminary summarisations before proceeding to carry out more sophisticated statistical analysis.

Although we believe that data analysis workshops of this kind are useful devices for teaching data analysis and statistics to users, we are not claiming it to be a substitute for conventional courses and consultancy sessions. It has its limitations and difficulties. However much we may try, not all types of statistical problems will arise in a few data sets. Issues such as design of experiments and organisation and conduct of surveys cannot be completely satisfactorily tackled in this type of programme.

Acknowledgement

I would like to thank Partha P Majumder for help on many counts: for moral support for this idea, for help in working out details of the workshop, for organising a second workshop along lines similar to the first and sharing its experiences, for the preparation of the proforma sent to intending participants, and for comments and suggestions on an earlier draft of this article.