Journal of the Korean Statistical Society (2001), 30: 2, pp 219–230

Patterns of Data Analysis?

Antony Unwin [1]

ABSTRACT

How do you carry out data analysis? There are few texts and little theory. One approach could be to use a pattern language, an idea which has been successful in fields as diverse as town planning and software engineering. Patterns for data analysis are defined and discussed, illustrated with examples.

Keywords: Data Analysis, EDA, Patterns

1. Introduction

Data Analysis and Statistics have an uneasy relationship with one another. For some there is no real need to differentiate between them, for others Data Analysis is a part of Statistics and for yet others Statistics is a part of Data Analysis (Huber[1997]). Tukey's distinction (Tukey [1962]), that statisticians operated more like judges, testing clearly specified hypotheses, while data analysts worked more like detectives, open to all sorts of different ideas, is extreme but helps set the scene. In particular, this view calls attention to the formal structures of Statistics compared to the undefined procedures of Data Analysis.

Statistics has much theory and can be taught as a subject in its own right without any reference to applications. It is hard to put together a theoretical structure for Data Analysis and it is impossible to imagine teaching it without applications. Data Analysis is more about methodology than technique (as indeed good statistical practice is too). Methodology is hard to teach and hard to write about. There are thousands of books on Statistics, including many excellent ones, but very few that describe Data Analysis - whatever their titles may claim. One of the rare examples is Tukey's famous EDA book (Tukey [1977]). In a sense it could not have been written at a worse time. Tukey described the exploratory nature of Data Analysis, but at a time when analyses had still to be carried out

[1]Department of Computer-Oriented Statistics and Data Analysis, University of Augsburg, 86135 Augsburg, Germany

by hand. Rather pessimistically, he wrote in the final chapter: "Now - and in the directly foreseeable future - there will be a place for hand calculation." Had he been writing but two or three years later, he would have been able to recommend personal computers, which even in those early years could still do far more than someone with pencil and paper. This is not to invalidate his ideas, but they would have been so different had he had the flexibility and speed of more modern computers at his disposal. He could certainly have discussed much larger data sets than the very small ones analysed in EDA. Data Analysis (and Statistics) is not limited to the tiny data sets often discussed in textbooks.

The lack of a theoretical methodology for Data Analysis is a handicap. Progress is made when formal structures are in place, which can be extended and developed. However, this lack of a methodology is not surprising. Methodological research in many fields brings problematic progress, if any. For example, the huge amount of effort put into the discussion of scientific methodology has led to many theoretical constructs, but not conspicuously to advances in practice.

One recent attractive approach in a related field is the study of patterns in software engineering. Instead of developing an all-encompassing general theory ("top-down"), an incremental bottom-up approach is followed (although the aim is to generate an overall framework). Certain task patterns arise frequently and so it is suggested that these are stored and organised in a uniform structure and format. They should be related to one another to form a pattern language. There is no claim made that all situations are covered, but standard and recommended solutions are available for commonly occurring ones. The initial development of the idea is commonly attributed to the architect Christopher Alexander and his group (Alexander et al [1977]), who were interested in town planning. They proposed different levels of patterns, which were closely interwoven to form a pattern language. It is the thesis of this paper that patterns may be a valuable and instructive way to view Data Analysis.

2. Structure of a Pattern

The term pattern can be used in different ways in English, but the meaning that most might think of initially is of a repeating design. The patterns we are talking about here have something of this, because of the requirement that they occur frequently, but it is more in the sense of a recommended pattern of behaviour in the face of a frequently occurring situation. We are not talking about patterns in results, such as the patterns it is suggested we look for in residual

plots from regressions. Textbooks on regression analysis usually include several plots of residuals against model predicted values. Illustrative plots are shown for cases where there are influential outliers, where the error variance is not constant, or where a transformation of the data might be carried out. Although Cook and Weisberg ([1999]) have pointed out that there may be other explanations for the forms seen in the displays, these will be rare and should only be considered if the context of the data suggests it. These displays are patterns of a kind, but not in our sense. They are usually idealistic and never seen in such a form in practice, but their principal features can be recognised and fairly clear recommendations as to what should be done can be made if such a pattern is found. These patterns in results are useful in interpreting analyses, but do not include the context of the problem (see the comment on Cook and Weisberg above) and do not explicitly refer to the problem to be solved. A pattern in our sense for this case would be whether a model is a good fit or not and looking at residual plots would be part of the solution.

Alexander says that "every pattern we define must be formulated in the form of a rule which establishes a relationship between a context, a system of forces which arises in that context, and a configuration, which allows these forces to resolve themselves in that context." There is no reference to a problem, rather to a context and its resolution. When computer scientists picked up the idea of a pattern language, they chose a more direct formulation and wrote of problems and solutions (for a readable and thorough introduction see Coplien [1996]). This is the approach adopted here.

There is no fixed standard for presenting patterns, but it is obviously necessary that descriptions include the problem, the context in which it arises, the system of forces operating, and the solution together with a rationale explaining why the solution works. Additional attributes such as a meaningful name, a visual representation, which evokes the pattern, and, examples of applications are valuable and relevant but not as central. However, it is important to list related patterns, whether associated, superior or subsidiary. It is the links between patterns, which convey the pattern language.

To be useful patterns have to be both general and flexible. Tightly defined contexts and restrictive assumptions will limit their applicability. In considering something to be a pattern it is instructive to consider the five properties of good patterns suggested by Coplien [1996]:

It solves a problem. Good patterns include solutions, they are not abstractions.

It is a proven concept. Good patterns have a track record in practice.

The solution isn't obvious. Good patterns are indirect, not from first principles.
It describes a relationship. Not just modules, but deeper system structures.
The pattern has a significant human component. Appeal to aesthetics and utility.

The first two properties emphasise the practical nature of patterns. They should genuinely solve frequently arising problems. The third property is more debateable. It is possibly attempting to suggest that there must be something innovative about the pattern. But what is innovative to one person is common sense to another so this is more an aesthetic criterion (which may also be useful in limiting the numbers of patterns proposed). The fourth property demands depth, but this is again a subjective issue. Examining data with an interactive histogram may be a simple idea, but can still reveal more than an "optimal" histogram calculated by some complex (and deep?) algorithm. The final property is somewhat ambiguous. Anything that can be automatically programmed is not a pattern. There may be much in a pattern, which should be programmed, but the whole is a support for human judgement and the human contribution is indispensable. A good pattern should be useful, but also attractive. It should seem like a good solution.

There would be a danger in following patterns rigidly. They are not supposed to be complete cookbook recipes with exact measurements and precise instructions. They are more like general principles, which can be stated qualitatively, but are not specified in technical detail. Each application requires the involvement of human judgement as well.

3. Higher-level Patterns of Data Analysis

3.1. Outliers

An interesting example of a general problem in Data Analysis is whether there are outliers in the data or not. Cases can be outliers in many different ways: they may be outliers on every variable, they may be outliers on a single variable, they may be multivariate outliers yet not extreme on any one variable. They may be outliers relative to the data set as a whole or only relative to a selected subset of interest. Outliers may be single cases or groups of cases. All these are data dependent properties of the outliers, which say nothing about why cases are outliers. They might be outliers because of errors in measurements or because of transcription errors. They might be extreme and unusual values, but nevertheless without error. They might be due to a mixing of distributions or due to a contamination. There are many possible explanations. How can all this

be translated into a pattern? One structure might be the following.

Name: Outliers

Problem: Are there outliers in the data set?

Context: Outliers may suggest quality deficiencies in a data set. The presence of outliers may invalidate or adversely affect the use of some statistical procedures. Outliers may indicate some cases of special interest (e.g. unusually high numbers of telephone calls may be due to fraud).

Forces: Difficulties in data collection may lead to poor quality data with errors. Imprecise sampling procedures may lead to contamination. Unknown background factors may lead to mixtures of distributions. Small sample sizes may suggest outliers, whereas large samples sizes would not. Identification of outliers depends on the standards used: relative to what are they outliers? Has the comparison been made with the proper population? (e.g. The patterns of telephone calls from a business will be different from those from a private number.)

Solution: Examine univariate and multivariate graphic displays interactively. Take account of available metainformation (e.g. numbers of calls cannot be negative). Discuss cases with domain experts. Check how the data were collected.

Examples: Telephone fraud. See also Barnett and Lewis [1997].

Resulting context:

Cases may be excluded, either as errors or because they are to be analysed separately as special cases. The remaining data points may then be analysed by standard methods. Robust methods may be used to downweight the effects of outliers, if they are kept in the data set.

Rationale: Identifying outliers and discussing them with data set owners brings greater understanding of the data set and more reliable interpretation of final results. (e.g. In a medical trial there may be additional information available on the patients.)

Related patterns:

Outliers on a single variable (the most common example when we think of outliers). Robust analysis methods. Smoothing.

Outliers is a high-level pattern, something which should be considered very

early on in an analysis. Specific methods for determining and explaining outliers depend more on the kinds of outlier that appear to be present. A lower level related example ("Single Variable Outliers") is given in the next section.

3.2. Comparison of proportions

Statistics is all about comparisons and one of the trickier areas is in comparing rates or proportions. This can be reasonably straightforward in a tightly designed study, but is far more complicated in a large survey. How many possible occupational groups are there? How many different ways of not answering a question? (No answer, no opinion, don't know,$\cdots$) Surveys often include so many questions anyway that one wonders if anyone can have filled them out with appropriate diligence. A pattern might be outlined in the following way.

Name: Comparing proportions

Problem: Do proportions in different groups differ?

Context: Categorical data arise commonly in surveys and opinion polls. Identification of meaningful differences between groups is of interest. Groups may be defined on one or more variables and determining key groups may be hard. Often there are too many categories for automatic statistical analysis and graphical exploration is valuable for picking out factors to be analysed in depth. Groups may have to be combined in a sensible fashion. The large numbers of questions posed in surveys complicates analysis.

Forces: Groups may be of different sizes, which makes direct comparison more difficult for the layman. There may be other groups which are not relevant for the comparison, but which confuse the picture. It can be appropriate to carry out a series of comparisons rather than a single one. Other factors may be influential on the rates and have to be kept in mind (cf Simpson's paradox).

Solution: Use spine plots rather than bar charts to display single variables, and mosaic plots for multivariate categorical displays. Incorporate confidence bands to approximately assess statistical significance. Check whether further factors might explain any effects observed better by using linking.

Examples: Did more treatment group patients die than control group patients?

Are promotion rates higher for men than for women after qualifications are taken into account? How does support for political parties differ between, say, old male rural voters and young female urban voters?

Resulting context:

Reduction of both the number of variables and the number of categories in variables allows more formal statistical modelling. Mosaic plots can be used to explore multivariate categorical relationships amongst a smaller number of variables.

Rationale: Exploratory comparisons are better made visually than numerically, when dealing with many groups or combinations of groups.

Related patterns:

Reorganising barcharts. Mosaic plots. Fluctuation diagrams. Contingency tables. Loglinear modelling. Logistic modelling.

Comparing proportions is a basic pattern, which will arise whenever there are categorical variables. Initial investigations will be concerned primarily with simple groups, while deeper analyses will incorporate more factors. Graphical comparisons are informative for the evaluation of the results of modelling to explore the sensitivity and relevance of the models produced.

4. Lower-level Data Analysis Patterns

The boundaries between patterns of different levels can be difficult to draw, but they are not intended to be procedures hewn in stone, rather suggestions which give analytic guidance. The two patterns described in this section are more specific than the two discussed already, but they are closely connected to them.

4.1. Outliers

Name: Single variable outliers

Problem: Is there an outlier in the distribution of a variable?

Context: An early stage of any analysis is to calculate statistics for each variable. Both mean and variance estimates are very sensitive to outliers. Standardising unchecked variables (e.g. before a cluster analysis) can seriously mislead.

Forces: One outlier may mask another, so that no outliers are identified.

One outlier may swamp another point so that too many are identified. The larger the data set, the more likely some data collection or transcription error has occurred. Real extreme values may be of more interest than the rest of the data in some applications.

Solution: Examine boxplots and histograms of the data. Query any outlier to see if a recording error is likely. Link to graphics for other variables to see if the case is an outlier on other dimensions.

Examples: Unusually long pregnancies may imply mistaken identification of father (see Barnett and Lewis [1997]). Age distribution in a survey (see below). Income distributions.

Resulting context:

Removing one outlier may suggest others. Further steps may depend on whether an explanation for the outlier has been found.

Rationale: Different visual displays reveal different aspects of the data. Boxplots highlight outliers, but give no indication of data set size. Histograms estimate the variable's density, but may not show outliers in a large data set due to insufficient display resolution.

Related patterns:

Outliers. Robust estimation.

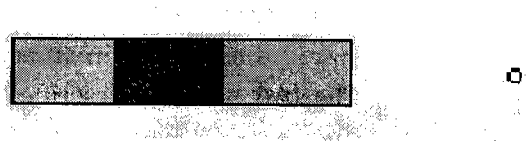

Figure 1 shows a boxplot of the age distribution of respondents in a survey (the Rochdale data set referred to in Whittaker [1990]). There is a lone outlier, but what action should be taken?

(a) query the point and the maximum of the rest of the distribution to see if the actual values give a clue as to what is going on. The exact value of the point might suggest a mistyping. The maximum of the rest might support the evidence of an error. (In the example the extreme value was 88 and all others were less than or equal to 65. Since the study was concerned with influences on whether women work or not and almost all women retire by age 65, the point is probably an error.)

(b) link to displays of other variables to see if the error can be confirmed or the extreme value given some support. (In this case there was a child less than 13 years old in the family, so the age of 88 was doubtful.)

(c) view other displays of the same variable. (For these data, a dotplot reveals the granular nature of the age data while the histogram in Figure 2 emphasises the large size of the data set, over 650 cases, making the outlier seem even more unusual. The size of the dataset is not apparent from the boxplot or the dotplot.)

All of these actions can be taken very quickly, but they will not necessarily be conclusive. Should more intensive investigation be thought worthwhile, then it would be necessary to check data collection procedures and to speak to people involved in the study.

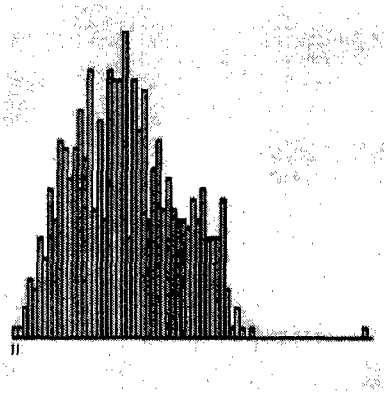

Figure 2 A histogram of the age distribution from figure 1.

4.2. Comparing proportions

Name: Reorganising barcharts

Problem: How can we make meaningful comparisons between bars?

Context: Barcharts compare absolute numbers, spineplots compare highlighted proportions. Barcharts (or spineplots) of many categories are difficult to interpret without structured ordering. The structure may usefully be data dependent (by count or by proportion highlighted) or depend on other available information (e.g. sort currencies by region). Categories may be clearly distinguishable (male/female) or possibly related (no opinion/don't know). Grouping related categories together may be beneficial.

Forces: Too many categories confuse and obscure. When there is a single large category, it is difficult to compare smaller categories as the scaling has to accommodate all. Differences between small categories will not be statistically significant and will rarely be actionable. Category definitions may be unclear (who is in "higher management"?). Data sources and quality of data collection may influence interpretation.

Solution: Use interactive switching between barcharts and spineplots. Use flexible reordering tools. Ordering by highlighting means that reordering can be by absolute numbers (barcharts) or by rates (spineplots). Highlighting can also be driven by selection on other variables. Group categories together which belong together.

Examples: A bank dealing in almost 50 currencies did over 90% of its business in only 5 currencies. Should the others be ignored or combined in making comparisons? (A reduced version of this data set is discussed below.) How do insurance claim rates vary by driver occupation?

Resulting context:

Understanding the one and two dimensional structure of categorical variables in a data set is a precursor to effective modelling.

Rationale: Categorical data need to be given structure to aid interpretation. Flexible reordering tools give many alternative views. Combining related categories generates groups which are large enough to be statistically compared.

Related patterns:

Comparing proportions. Loglinear modelling. Mosaic plots.

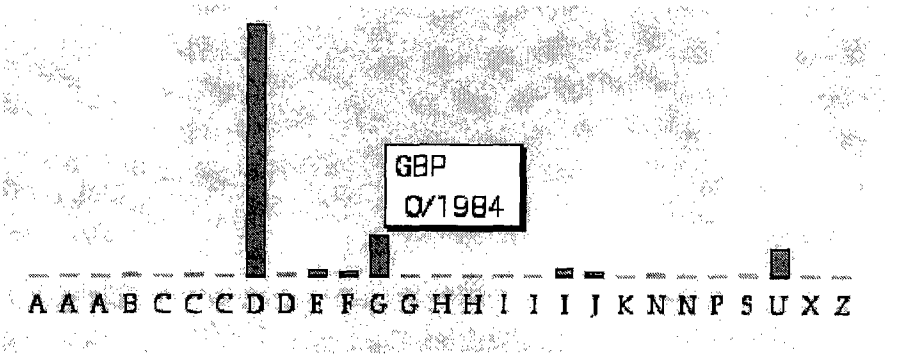

Figure 3 shows a barchart of deals by currency for a major bank. There are 27 different currencies represented, but 16 of them are each less than 0.5

After sorting by size and then grouping the smallest columns together, the graph in figure 2 is obtained. This is much clearer and it is now possible to read the currency labels directly. Experimenting with different orderings and with different groupings and using background information ensures that little distortion takes place and that no valuable information is lost.

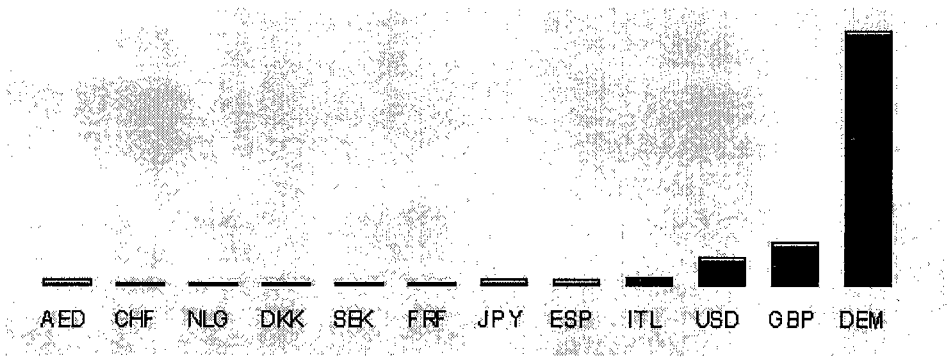

Figure 4 Bar chart of currencies by numbers of deals after sorting by count and then grouping together the smallest 16 into one new category on the left.

5. Conclusions and outlook

It is difficult to describe the key processes of good Data Analysis. Patterns and a pattern language offer one promising approach, though others may have potential too. The important thing is to have a methodology that may be described and taught, and which works in practice. Data Analysis tends to involve following up several ideas in parallel rather than trying to find a single optimum. Its methodology should reflect this and not be overly restrictive. All patterns are to some extent idealistic, but the treatment of the idealistic case indicates how the less clear cases may be dealt with. Human judgement is an essential component, but the patterns provide a framework within which judgement can be consistently employed.

Developing patterns is an incremental process. The success of the idea will depend on an increasing number of patterns being written and on a consensus pattern language emerging. This is a development in which all can participate and I look forward to many new insights into successful Data Analysis. The concept of patterns is a much deeper subject than can be effectively handled in

the space available here, but it is a thought-provoking way of regarding Data Analysis. Whether we adopt the approach or not, its discussion sheds light on a complex process that plays a central role in statistical practice.

Acknowledgement

Some of the ideas in this paper have been discussed with members and former members of the Augsburg group (Heike Hofmann, Stephan Lauer, Adi Wilhelm). I am grateful to them for their constructive criticism.

REFERENCES

Alexander, C., Ishikawa, S. and Silverstein, M. (1977). *A Pattern Language : Towns, Buildings, Construction.* Oxford University Press.

Barnett, V. and Lewis, T. (1997). *Outliers in Statistical Data (3rd ed.)*, Wiley.

Cook, R. D. and Weisberg, S. (1999). Graphs in Statistical Analysis: Is the Medium the Message?, *American Statistician*, **53**(1), 29-37.

Coplien, J. (1996). *Software Patterns*, New York: SIGS Books.

Huber, P. J. (1997). Speculations on the Path of Statistics. In Brillinger, D. R., Fernholz, L.T., Morgenthaler S. (Eds.), *The Practice of Data Analysis* (pp. 175-191). Princeton University Press.

Tukey, J. W. (1962). The future of data analysis, *Ann Math Stat*, **33**, 1-67

Tukey, J. W. (1977). *Exploratory Data Analysis*, London: Addison-Wesley.

Unwin, A. R. (1999). Requirements for interactive graphics software for exploratory data analysis, *Computational Statistics*, **14**, 7-22.

Velleman, P. F. (1997). The Philosophical Past and the Digital Future. In Brillinger, D.R., Fernholz, L.T., Morgenthaler S. (Eds.), *The Practice of Data Analysis.* Princeton University Press.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester: Wiley.