A DATA ANALYSIS TOOL THAT ORGANIZES ANALYSIS BY VARIABLE TYPES ®

Rodney Carr
Deakin University
Australia

XLStatistics is a set of Excel workbooks for analysis of data that has the various analysis tools and methods organized according to the number and type of variables involved. Most introductory courses in statistics start out with a discussion of the different types of variables, but very few data analysis packages are organized along these lines. This can make it difficult for students, and may be a contributing reason for the common "What test do I use?" question. We describe the XLStatistics package and show how it may help to overcome some of the common problems encountered by students.

INTRODUCTION

"What test do I use?" This is maybe the most common question asked by students learning statistics. We suggest that this question, and ones like it, may be avoided to a large extent by organising the tools and methods of data analysis for them in a simple way. Most instructors point out to students that the type of analyses they can employ are determined to a very large extent by the number and type of variables involved. Essentially all textbooks use the same idea to help organize the various topics. It is a very important guiding principle – the appropriate analysis techniques to use in any given investigation are determined to a large extent by the number and type of variables involved. However, although many courses and texts start out this way, later on many switch to an emphasis on population parameters, or talk in terms of the numbers of populations involved, and thus to a large extent destroy the simplicity of the variable-driven approach. Most of the standard statistical analysis software packages do not organize the tools according to the number and type of variables involved either. We will present a brief discussion of the issues and present a package (XLStatistics) that does have the various analyses organised according to the number and type of variables.

LEVELS OF MEASUREMENT AND APPROPRIATE ANALYSES

Measured variables can be classified in many ways. One standard way follows Stevens (1946) and classifies variables into Nominal, Ordinal, Interval or Ratio. There has been much debate and controversy for over 50 years about this classification. Sarle (1995) and Velleman and Wilkinson (1993) have a good discussion of the issues. And although the Stevens classification can be useful in many problems (for example, in scale development), for analysis, from a pragmatic point of view, the methods that are appropriate mostly depend on the simpler classification of variables into Categorical or Numerical. For a single Categorical variable, for example, appropriate analyses include various styles of tables showing frequencies or proportions, bar charts of frequencies or proportions, pie charts for proportions, tests for proportions and goodness-of-fit tests. Analyses for a single Numerical variable include frequency and relative frequency histograms, measures of location and spread and associates tests on these quantities (t-tests, etc). The idea extends – if there are more variables the appropriate analysis methods depend to a large extent on the number and type of variables involved. For example, if we have an investigation involving 2 Numerical variables, the appropriate analyses can involve scatterplots, fitting of curves, correlation or regression, etc. If we have an investigation involving 1 Numerical variable and 2 Categorical variables, we might use a pivot table showing summaries of the numerical variable and carry out a 2-way ANOVA. The point is that the relevant tools and techniques depend to a large extent on the number and type of variables involved. This is an important idea - we will call it the "Num/Cat principle".

Actually, this in not so cut-and-dry as it might first appear because often variables can be classified in different ways. For example, variables of the intermediate types in the Stevens classification (Ordinal and Interval) can often be treated as either Categorical or Numerical. Luckily (with appropriate minor adjustments) appropriate results are usually obtained no matter

how such variables are treated. For example, consider the data shown below, obtained from a 7-point Likert-scale question on a questionnaire:

1, 5, 2, 3, 3, 3, 4, 3, 2, 2, 3, 6, 5, 5, 5, 0, 3, 2, 3, 6, 4, 5, 4, 3, 5, 6

This data is probably best classified as values for an ordinal variable (a categorical variable with an associated order for the categories), and if treated this way we might summarize the data using a bar chart (Figure 1):

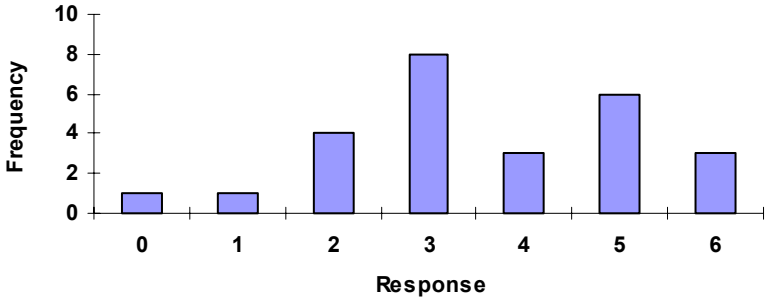

Figure 1. Bar Chart for Categorical Data.

This is very similar to the standard bar chart that might be drawn for a Nominal variable (a categorical variable where there is no natural order associated with the categories), but the various categories are shown in the appropriate order. The chart also shows the number rather than labels for the various categories, but labels such as "Strongly disagree", "Disagree", etc, could be used if it were important to show this information. If we treat the variable as a numerical variable, we might summarize the data using a bar chart like that shown Figure 2.

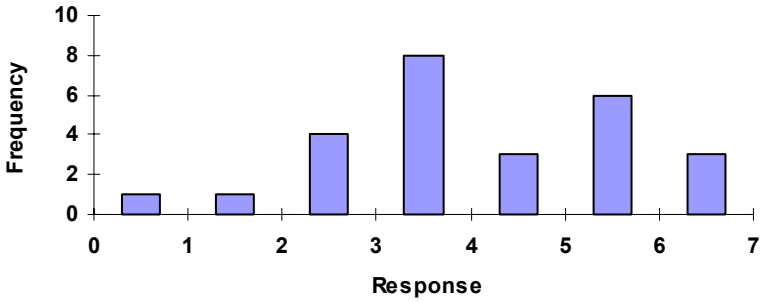

Figure 2. Bar Chart for Numerical Data.

The main difference between the charts in Figures 1 and 2 is the way we draw the horizontal axis. In Figure 1, the labels mark the actual classes; in Figure 2 we have labelled the endpoints (in the chart, all bar the rightmost class are closed at the lower end; the last class is closed both ends). Marking endpoints of classes tends to give a better feeling that the variable changes continuously. To go further, if we really want to convey the idea that the numbers feel that the numbers are varying continuously, we might draw a frequency histogram, as in Figure 3:

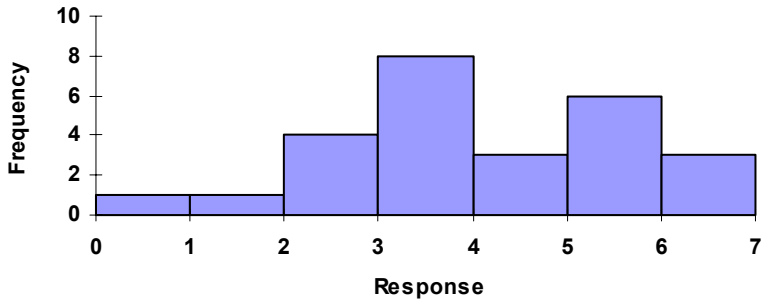

Figure 3. Frequency Histogram for Numerical Data.

All the above charts can be appropriate for Ordinal data, depending on the situation. If there are only a few possible responses on a Likert-scale question, for example, a chart such as Figure 1 might be preferred, but if respondents are tempted to tick in between labels such as "Strongly disagee" and "Disagree", a chart such as in Figures 2 or 3 might be better. But none are usually actually inappropriate, and much depends on what the analyst wants to highlight.

This ambiguity in classification of variables extends to other analyses, too. For example, with the above data, analysts might sometimes use a pivot table showing percentages to summarise the data and might carry out a test on the proportions (which is treating the variable as categorical). Or they might treat the variable as numerical and maybe use the median or the mean as a summary on the average value and then do a t-test. The point is that none of these analyses are really inappropriate, and choices depend to a large extent on personal taste and the situation.

The ambiguity is also not limited to variables of Ordinal type. For example, variables that are quite clearly of Ratio type (where it makes sense to compare values by taking ratios) are often "Categorized" (maybe into "Low", "Medium" and "High") and then treated as categorical. This also occurs when a frequency histogram of a Ratio variable is drawn – the values are grouped into classes, thus making the variable into a categorical variable.

HOW STUDENTS MANAGE VARIABLE CLASSIFICATION

In order to use the Num/Cat principle (the idea that analysis methods depend on the number and type of variables), we must be able to classify variables. And, in spite of all the ambiguity in how variables can be treated, students seem to have no trouble classifying them into the broad categories, Numerical or Categorical. Classroom observations indicate that they do not even seem to worry too much when Nominal variables are coded using numbers, such as using "0" and "1" to stand for "Female" and "Male". Ambiguity seems to not cause problems, either - they seem to be not too worried by our being able to treat some variables as either Categorical or Numerical, though evidence from actual class observations suggests that they usually opt to treat Ordinal variables as Numerical (which is usually fine).

The greatest difficulty that students have when determining the number and type of variables involved arises when data is presented in "unstacked" form, such as sets of measurements of a single numerical quantity, but from different populations. However, most statistical packages require their data to be in stacked form, and, usually at the same time as variable-classification is discussed, students are trained to stack their data, ready for analysis. (There are a few exceptions to this "data-must-be-stacked" rule. Minitab, for example, allows users to carry out a one-way AVOVA with data in unstacked form.) This causes some difficulties for students, but it is not usually serious. And because it is normally done near the start of a course, there is usually enough time for students to get used to the idea and to carry out a thorough training. So, it appears that students learn to identify the different variables in data without too many problems.

PUSHING THE NUM/CAT PRINCIPLE - XLStatistics

As mentioned in the introduction, most courses and texts start out with a discussion of different types of variables. Most also have separate topics presenting analysis tools for data with a single numerical variable, for a single categorical variable, and for 2 numerical variables.

However, the structure is often not followed through to the end. For example, it is quite common for texts to talk about numerical data from two or more populations, without specifically pointing out that the resulting data set, when appropriately organised for analysis, has one numerical variable (the numerical data) and one categorical variable (comprising labels for the populations). Titles such as "Comparing two or more populations" are used instead of relying on the classification by number and type of variables. In fact, in this situation there is often a confusion where measurements that belong to the different populations are treated as separate variables, rather than using the categorical variable to distinguish between the various populations. This may be even more confusing to students when they are required to stack the data for analysis. The overall effect may be that the structure of the course (or flowchart) becomes more complicated than is necessary.

The practice of not organising the work according to the number and type of variables can also be found in most of the common data analysis packages. SPSS, for example, has separate menus for graphs and other analyses, and does not organise the items on these menus in a manner that guides the user to an appropriate analysis according to the different types of variables involved. (Interestingly, the Statistics Coach in SPSS has been designed to guide users through to appropriate analyses, and depends on users classifying variables.) So before using a package like SPSS, students must normally first know which are the analyses that are appropriate for their type of data, then where to find them in the menu. This is, of course, possible to do, but it is another step in the learning process that can be eliminated. We suggest that, to a large extent, it is the cause of the "What test do I use?" question.

Given the above discussion, it should be apparent that there are two things that are needed in order to retain an emphasis on the number and type of variables all the way through a course:

1. A text that is presented this way;
2. A software package that is organised this way.

There is at least one text that is organised this way ("Practical Statistics", available from the author), but for the remainder of this paper I wish to discuss aspects of XLStatistics, a software package that has the analysis tools organised according to the Num/Cat principle.

XLStatistics is a set of about 80 Microsoft Excel workbooks that each contains separate or related analyses. Together the workbooks form a reasonably sophisticated data analysis package. The package contains most of the standard analysis tools, up to, for example, multiple regression. It does not at present carry out multivariate analyses such as PCA or cluster analysis, but modules are being added fairly regularly.

With 80 separate workbooks the package might sound quite cumbersome. However, the whole package is organised around the Num/Cat principle and in general students find it very easy to navigate through. There is only one workbook users need to open directly – XLStats.xls. It is the launch-pad and all the other workbooks are linked to this, either directly or indirectly. When opened, the launch-pad appears as shown in Figure 4.

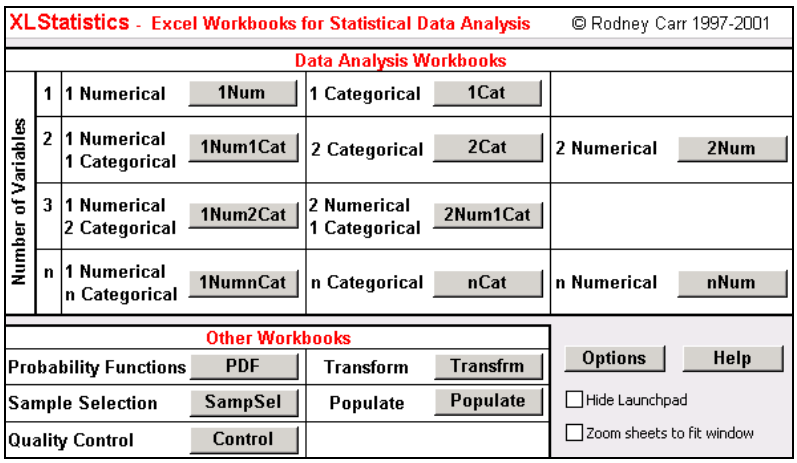

Figure 4. The XLStatistics Launchpad.

The buttons in the Data Analysis Workbooks area on this screen correspond to different combinations of variables. When one of these buttons is clicked, another Excel workbook opens that contains tools appropriate for the corresponding combination of variables. The name of the workbook that is opened is actually the same as the label on the button, so that, for example, clicking the button labelled 1Num1Cat will open the workbook named 1Num1Cat, used for analysing data with 1 Numerical variable and 1Categorical variable. At the same time as the launchpad appears, an XLStatistics menu appears on Excel's main menu bar, and all the workbooks can be selected from there – this is how the selections are normally made, in fact.

To illustrate how XLStatistics is used, suppose that students are asked to determine an estimate of the average score on the Likert-scale question with data as presented previously. Students organize the data, and recognize it as Numerical data. They then highlight the data and select "1Num" from the XLStatistics menu. This opens 1Num and the data is entered into the workbook's Data area. The students are then presented with analyses appropriate for this type of data. In particular, the frequency histograms shown previously appear and a table of numerical summaries is given. An image of part of the screen is shown in Figure 5.

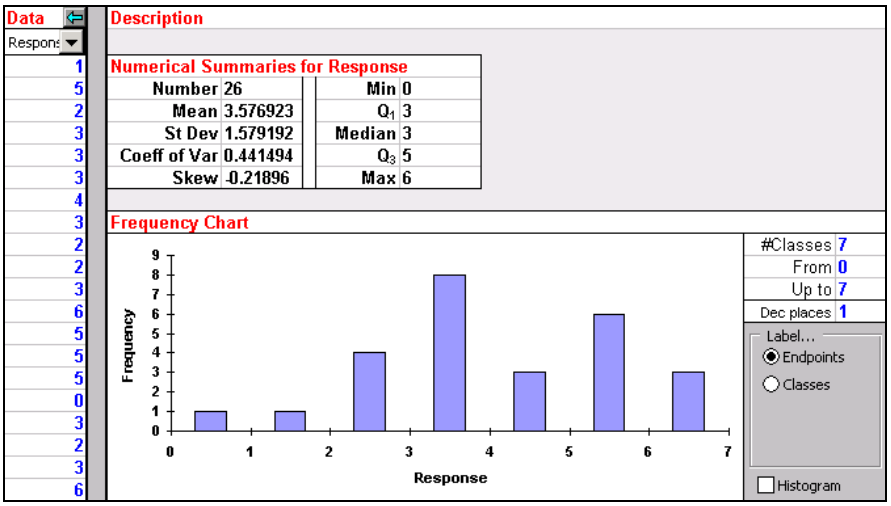

Figure 5. Part of the "Data & Description" Sheet in 1Num.

The student must choose a suitable measure of "average". (It is the instructor's task to explain what the various analyses are for, and how to interpret and use them in an appropriate manner. However, much of this can be achieved within XLStatistics itself by simply varying the data or other information to illustrate the various features and properties.) Looking at the above frequency chart, it seems that the mean is an appropriate measure to use. If the data were randomly-chosen, the student could then click on the "Tests" sheet, where they find a corresponding confidence interval for the population mean, as pictured in Figure 6.

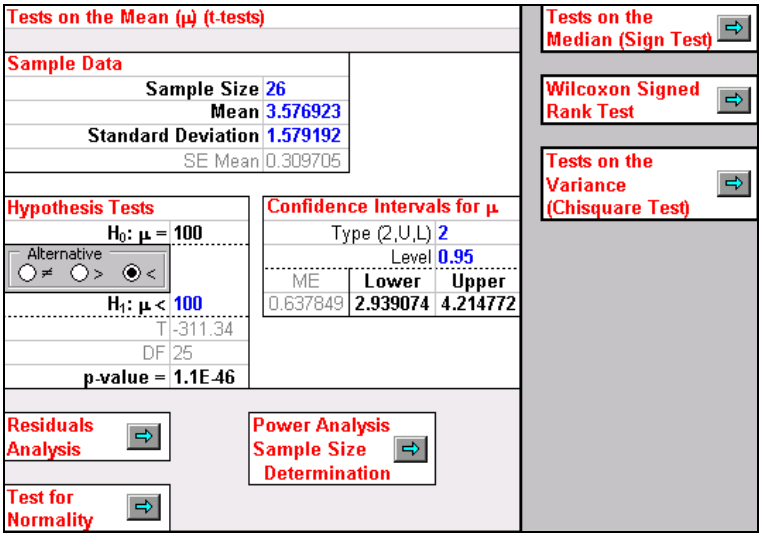

Figure 6. The "Tests" Sheet in 1Num.

(Again, it is the instructor's task to explain what the confidence interval is, and show the relationship between the confidence level and the margin of error, etc, but much of this work can be achieved interactively within XLStatistics itself.) The final report would state that the true population mean is unknown, but that it most likely lies between about 2.9 and 4.2.

Notice that students do not need to know the name of the test to use to carry out the above analysis – they are lead to the appropriate work once they have identified that they are dealing with 1 Numerical variable. (However, the name of the test is given, so, if necessary, they can look up the details in a textbook, or use the name in a report.) Notice also that it would be easy for them to choose a test for the median, possibly more appropriate if the data had been skewed, for example.

The simple idea of organising analyses according to the type of variables involved seems to be enough to allow students to get started with their analyses in a simpler way that when they use other packages. Some direct evidence of this comes from one recent introductory business statistics course. The course had a mixture of on- and off-campus students. Students were given the choice of using either XLStatistics or another package that has a more "standard" structure. Out of the 187 students who submitted the assignment (47 on-campus and 121 off-campus), over 75% of the used XLStatistics. When asked why they choose to use XLStatistics there were about 20 responses, most stating things like "I used more of XLStatistics because I found that it was extremely easy to use." The most interesting response was from a student who was repeating the unit: "XLStatistics seems to be easy to understand and I think for the first time I have been able to understand what is required, how come stats is important and how useful it can be for business." Of course, this is not conclusive evidence that using XLStatistics does produce better learning, but it is suggestive. Certainly, further investigation is warranted.

REFERENCES

Stevens, S.S. (1946). On the theory of scales and measurement. *Science*, 103, 667-680.

Searle, W.S. (1995). *Disseminations of the International Statistical Applications Institute* (4th edn.) (pp. 61-66). Wichita: ACG Press. See also <ftp://ftp.sas.com/pub/neural/measurement.html>

Velleman, P., & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician*, 47:1, 65-72. See also <http://www.spss.com/research/wilkinson/Publications/Stevens.pdf>