SMARTER STATISTICS – DON'T GO DOWN WITH THE TITANIC

Andrew Stewart

Presbyterian Ladies' College

The implementation of new technologies can move the teaching of statistics in Years 7 – 12 beyond repetitive calculation and display of descriptive statistics. Drawing on resources provided and developed during participation in the STATSMART project, suggestions are offered for improving statistics teaching and learning.

Introduction

Statistical analysis developed dramatically in the later years of the nineteenth century and the early years of the twentieth century. Karl Pearson, Professor of Applied Mathematics at the University College, London, developed mathematical methods for studying the processes of heredity and evolution. Between 1893 and 1912 he published 18 papers entitled Mathematical Contributions to the Theory of Evolution which contained his valuable contributions to statistics, including the correlation coefficient which now bears his name. Pearson used large samples from which he tried to deduce correlations (HREF1). Meanwhile another great statistical titan, Ronald Fisher, used small samples and tried to deduce causes (HREF2). These two strong advocates of statistical methods had a long, bitter, and very public dispute until Pearson's death in 1936.

Both men, while giving a lot of their time to their area of expertise, could also be quite narrow minded on some issues and did not suffer fools greatly.

In some respects, this narrow-minded approach to statistics is apparent in what many of us are doing in the classroom. Textbooks for years 7 to 10 show that the Statistics

chapters contain large sets of numbers for numerical analysis – an activity which soon loses the interest of many students.

So what can we do about it?

From my involvement with the Statsmart research project – a project looking at how well students are learning their Statistics over a number of years – there are two changes that I suggest we can make to our teaching that will make a difference.

Suggested Changes

First Change

The first suggested change involves visual presentation of the data. This generation of students is very visual – they respond better to visual material than reading. We can make Statistics more visually appealing through the careful use of one of a number of technologies – software or hardware or a combination of both. These technologies also enable us to analyse much larger data sets than we would have been able to by hand or with simple calculators.

The software approach involves programs such as a spreadsheet (for example Excel) or purpose designed statistical analysis software (for example TinkerPlots or Fathom). Visual presentation of data in a dot plot or a scatterplot, or analysis in a boxplot makes it easier for students to begin to understand or interpret what is going on.

Scatterplots, histograms and boxplots can be assembled in spreadsheets. The use of spreadsheet templates would enable the rapid presentation and analysis of a number of data sets, and then time could be spent on discussing the results of these analyses.

Specialist software such as TinkerPlots has been designed to make the data collection and analysis easy, so that more time can be spent in discussing the results obtained, and perhaps asking more questions about the data and analytical results.

The hardware approach involves graphic calculators that contain functions or programs which allow the display and analysis of statistical data. Examples of these kinds of calculators include the TI-84, TI-Nspire or the Casio ClassPad. While presenting information on a much smaller screen than a computer, they offer more portability and easier access in many cases.

Second Change

The second suggested change involves presentation of data in context. Rather than focusing on "assembled" data sets with neat answers, use a context to help explain why particular measures are used.

Given the right tools and contexts, students are capable of quite powerful statistical analysis. Looking for relationships between data variables does not necessarily mean that regression tools need to be used – students at Year 7 are quite capable of determining whether a relationship exists between two variables from examining a scatterplot!

There is a large amount of real world data that can be drawn upon for class activities. I have used a large number of sport-based data sets, particularly in teaching Further Mathematics at Year 12. Until recently, data from female sports has been hard to find, and thus male sports such as AFL football or cricket have acted as the main source. The growth in women's sport, and the growing space it receives in the media, means that more data is becoming available. Other aspects such as health issues, consumer products or the key constituents of foods can also be analysed statistically. Sources of data could include accurately maintained web pages, publications (books, periodicals, newspapers) or personal collection.

Bringing Context and Visualisation together

Teaching Basic Concepts

One of the most powerful visual demonstrations I use in a classroom shows how a boxplot is drawn. On a dotplot where the values are shown distinctly separated from each other, a vertical line is drawn at the median value. The quartile values are shown by vertical lines (each quartile being a median of half the data), and finally the maximum and minimum values are shown with vertical lines. The two quartile lines form each end of the box, and lines are drawn from the box to the minimum and maximum values. This activity can be demonstrated to the class, or carried out by students on computer or a printed copy of the dotplot. In TinkerPlots, I use the program to plot a boxplot on the data and it comes down on top of the one that I have drawn, as shown in Figure 1 on the following page.

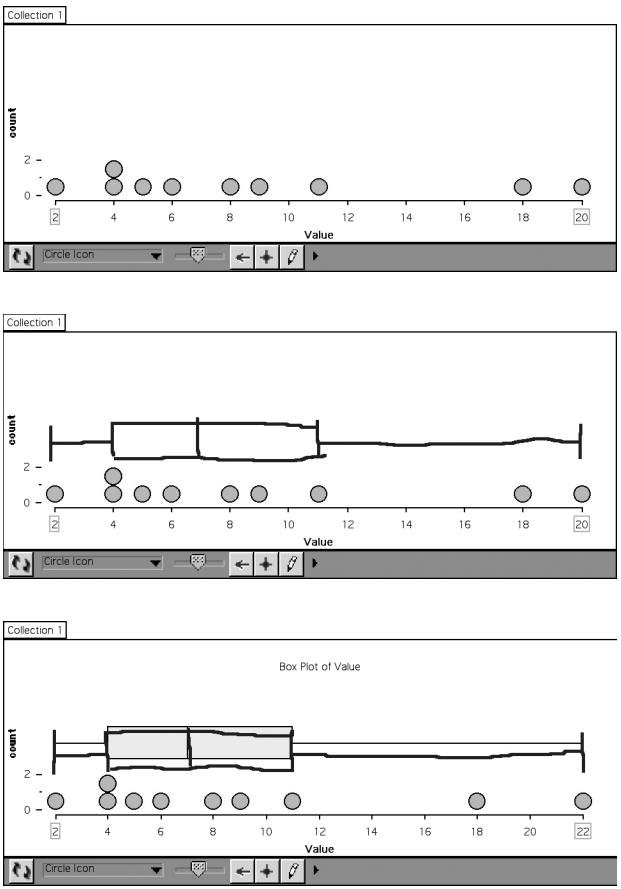

Figure 1. Key steps in demonstrating boxplot construction

Illustrating what happens when an outlier is present follows the same procedure up to the drawing of the box. From the quartile values we can determine the Inter Quartile Range and hence calculate the whisker limit – it is drawn to the nearest data point whose value is distant by up to 1.5 times the Interquartile Range from either of the quartile values. Points which lie beyond these limits are highlighted as outliers as shown in Figure 2.

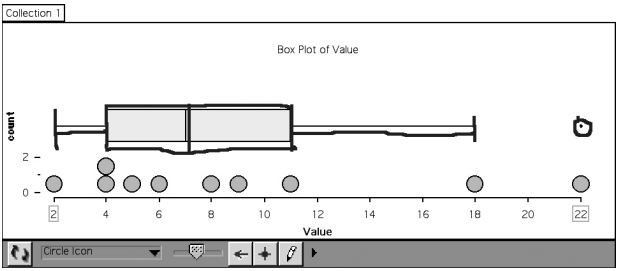

Figure 2. Demonstrating outliers for a boxplot.

Visualisation is also very powerful in exploring the effect of outliers on the mean and median values of a data set. For some time I have used "assembled" data sets in which I have changed one value several times by a different order of magnitude each time and looked at what happens to the mean and median values. However I think it would make more impact if it were done with real data.

Jane Watson, at the University of Tasmania, is developing an Australian resource to support TinkerPlots. She has assembled a data set of the salt content of common household foods, with soy sauce as a conspicuous outlier. TinkerPlots allows you to hide one or more data values and be able to observe the effects on the mean and median values of the remaining data set. In a spreadsheet or in a calculator, we can replicate this by deleting the value(s) of particular cell(s). On a printed sheet, we could give a table of values as well as the dotplot of the data. This kind of data enables us to explore the ideas of which measure of centre is best for particular data sets.

Another useful exploration of mean versus median could come from analysing the weekend house sales results. The analyses published by the Geography Department at Monash University always quote median house prices. For example, on a relatively quiet weekend, the sale of one or two very highly priced properties can have a major effect on the mean as calculated. On Monday 10th August, the top price listed was $7.3 million, ahead of a number at about $2 million, with the lowest at about $250 000.

Challenging Investigations

Sir Donald Bradman's official batting average after each innings of his first ten innings is given in the table below, together with two of those individual innings scores (data sourced from HREF3). Cricket batting averages are calculated by dividing the sum of all

runs made by the number of innings in which the batsman was dismissed. (Clearly having a lot of not out scores will enable a batsman to have a high average score). In this table, ** indicates a "not out" score. A graph of these averages is shown underneath.

Innings No.	Scores	Average
1	18	18.00
2		9.50
3		32.67
4		52.50
5		50.00
6		51.33
7		61.57
8	37**	66.86
9		59.50
10		67.44

Table 1. Data table for Bradman scoring problem

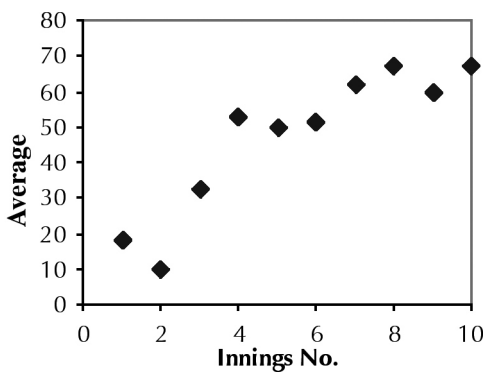

Figure 3. Graph of Bradman's batting average for his first ten innings

This activity turns the traditional average calculation process on its head, but enables us to have a discussion on what kind of values will cause an average to increase or decrease from its previous value. This activity will require careful calculation by students, and could be made more challenging by removing the value of the eighth innings, but leave the marker that it is "not out".

As an extra afterwards, we could look at the ten values, and discuss whether an "average" or a median is the best descriptor of the central location or typical score.

Technological advancement means that analysis of large data sets is now able to be carried out in schools. While a specialised package like TinkerPlots makes this quite easy, spreadsheets are a viable alternative that enable students to work for themselves or within prepared templates. With graphic calculators able to share programs or data files, many students can work on a common data set. All these technologies enable us in some way to selectively filter the data to look for particular subsets of data. Being able to "mine" data sets is a valuable skill in our information-rich society.

The Australian Bureau of Statistics CensusAtSchools project collected a large amount of non-personal data from hundreds of school-age students throughout Australia. Samples from this large dataset can be downloaded for analysis by students. This analysis can be univariate (mean, median or mode) looking at the centre of various data sets or a simple bivariate analysis (via scatterplot) where students look for relationships between two variables.

My colleagues teaching Year 7 at PLC have taken this concept one step further. Each of the one hundred plus students in Year 7 entered the answers to about thirty data questions into a large spreadsheet. Each student was then given a copy of the completed spreadsheet and asked to perform a series of analyses using TinkerPlots. There were a number of specific data sets that were to be examined and then the students were given free choice for other analyses. Teachers restricted options for data collection for students by, for example, limiting eye colour options to blue, brown or green. While students were advised of the units of measurement to be used, some data was either measured incorrectly (metres instead of centimetres), or entered incorrectly into the data set. As with the ABS CensusAtSchools data set, errors were not corrected prior to student access. The presence of "dirty" data leads to interesting discussions about the nature of errors and what could be done about them in the analysis.

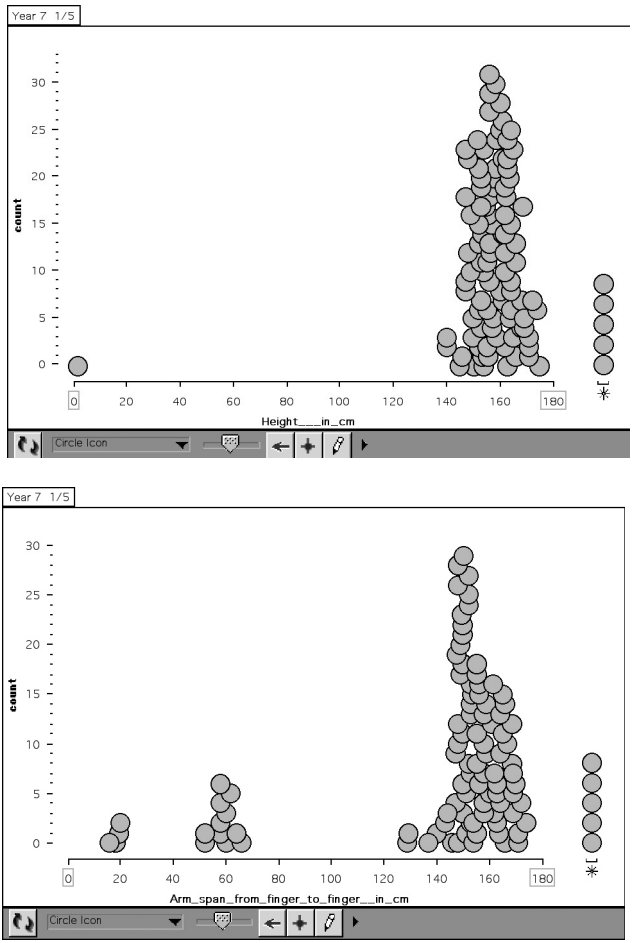

Figure 4. First plots of student data for height (above) and arm span (below).

In the uppermost plot in Figure 4, a student entered her height as 1.54 cm instead of 154 cm. In the lower plot, the data points at around 60 cm are most likely measurements made in inches instead of cms, and the values at about 20 cm are most likely hand span measurements instead of arm span measurements. In both diagrams, the stack of points at the extreme right hand side indicates that five students did not supply a value for this measurement.

At Year 8 at PLC we have extended this concept of analysing a large data set. The RMS Titanic, which sank in 1912 with a large loss of life, had about 2 500 people on board. A number of Internet sites have been established containing datasets about the Titanic passengers and crew. From one of these (HREF4), we set up a dataset with about a dozen items on each person on board, leading to a total of about 30 000 pieces of information. The data is very "dirty" as there are many gaps and inconsistencies. However, using TinkerPlots and its filter function, students have been able to explore and analyse the data quite effectively.

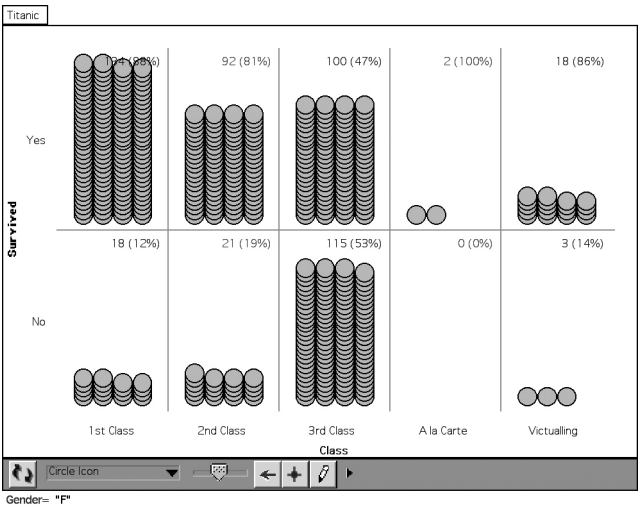

Figure 5. The graph shows the number (and percentage by class) of women who survived or died, together with their passenger class or crew area.

Figure 5 shows a two-way frequency table showing the traveling status of female passengers or the work areas of female crew and whether they survived or not. When first plotted by students, it provoked interesting discussion, as it showed that there were only 23 female crewmembers on the Titanic, and all but three survived. It also shows that over 80% of female passengers in first and second class survived, but only 47% of third class female passengers survived.

The Titanic story and data set will form the basis of a cross-curriculum activity for all our Year 8 students this year. While mining the data set last year, students found

that passengers came from many countries, with a large contingent boarding at Ireland and France prior to the Titanic heading out across the Atlantic. Thus a LOTE activity can be generated as well as ones for Geography (mapping the journey to its final resting place), History (gender balance in the crew), Physical Education and/or Science (effects of cold water on the human body) and English (looking at various media presentations of this event).

Conclusion

Statistics can be made more accessible and relevant to students by using data in context with a story, and using visualisation of the data in graph form to assist students in making sense of the data and carrying through its analysis. This works as well in teaching the basic concepts as in carrying through an investigation which reinforces the basic concepts.

References

Websites

HREF1: http://www-gap.dcs.st-and.ac.uk/~history/Biographies/Pearson.html
Karl Pearson Biography, University of St. Andrews, Scotland

HREF2: http://www-gap.dcs.st-and.ac.uk/~history/Biographies/Fisher.html
Ronald Fisher Biography, University of St. Andrews, Scotland

HREF3: http://www.cricinfo.com/ci/content/stats/index.html
Cricket Statistics, ESPNCricinfo

HREF4: http://www.encyclopedia-titanica.org/titanic_passenger_list/
Titanic Passenger data, Public interest site