What is ‘typical’ for different kinds of data?

Examples from the Melbourne Cup

Jane Watson
University of Tasmania
jane.watson@utas.edu.au

There are five words that are critical to an appreciation of what is ‘typical’ in data sets that students encounter across the middle school years. The first three words—mean, median and mode—are nouns that define measures of typicality in data sets. The next two words—categorical and numerical—are adjectives that describe the types of data sets to which we may wish to apply the measures.

The purpose of this article is to present and discuss in a down-to-earth fashion, with authentic data sets, the relationship of the three measures of typicality and the two types of data introduced in the curriculum. Although the mean, median and mode have been in mathematics text books for many years it is sobering to read in the research literature that teachers’ understanding of the measures is little different from that of students and both struggle (Jacobbe & Carvalho, 2011).

What does the Australian Curriculum say?

Mean, median, and mode are not explicitly mentioned in the *Australian Curriculum: Mathematics* (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2013) until Year 7, where we find the following:

> Calculate mean, median, mode and range for sets of data. Interpret these statistics in the context of data (ACMSP171)
> Describe and interpret data displays using median, mean and range (ACMSP172)

The inclusion of ‘context of data’ assumes an appreciation of which types of measures are associated with which contexts. Looking back through the curriculum we find that categorical data or variables are mentioned in Year 2, Year 3 and Year 5, whereas numerical data or variables are first mentioned in Year 5. This progression must reflect the writers’ appreciation that counting frequencies and comparing them across categories (e.g., favourite fruit) is achievable in these early years. It would have been useful,

[1] This article, based on data to 2012, was written for Top Drawer Teachers (http://topdrawer.aamt.edu.au) and is published here with the permission of Education Services Australia. © Education Services Australia. The plots in this article were created using the software *TinkerPlots* (Konold & Miller, 2011).

however, to introduce the term 'mode' when giving the example of 'most popular breakfast cereal' in Year 3, or when expecting students to interpret data displays (ACMSP070). The mode as 'most popular' or 'most frequent' is not a complex idea and could be seen as a way of describing what is typical for categorical data sets.

By Year 7, where mean, median, and mode are introduced, numerical data are the focus and although the mode is mentioned when students are asked to calculate the statistics and interpret in the context of data, it is not mentioned when students are asked to describe and interpret data displays. Once numerical data become the focus of statistical investigations, with the potential to employ the mean and median, it may not be surprising that the interest in the mode lessens. The emphasis on categorical data throughout the primary years, however, suggests that there may be confusion on the part of students about whether they can apply all of these new measures to the data sets with which they are most familiar.

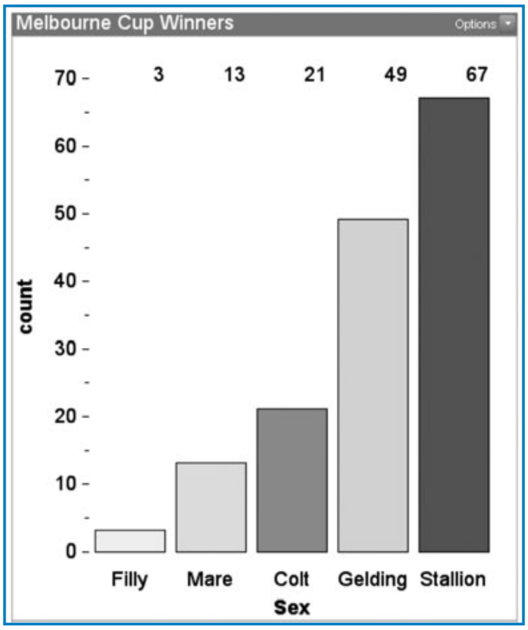

Figure 1. Sex of Melbourne Cup winners.

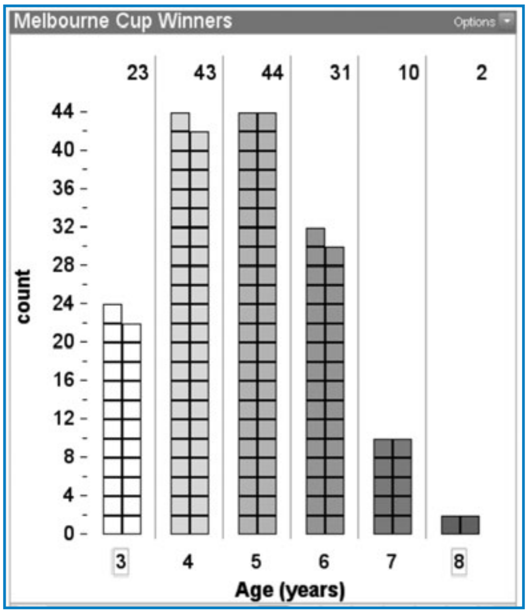

Figure 2. Age of Melbourne Cup winners.

How do we know which measure to use for which type of data?

The examples provided here are based on data collected for winners of the Melbourne Cup (Watson, Beswick, Brown, Callingham, Muir & Wright, 2011). In this Australian context, variables are available that are both categorical and numerical. If we think about the curriculum and start with categorical data, the measure of typical we need is the mode, which tells us which category or categories has/have the most values of the variable in it/them. In Figure 1 are five different sexes for the winners of the Melbourne Cup: Filly, Mare, Colt, Gelding, and Stallion.[2] We see from the column graph that the most common sex of winners of the Melbourne Cup is Stallion. We could say this is the modal or typical sex of winners.

Other sets made up of numerical rather than categorical values can also have modes. For example in looking at the Age of winners of the Melbourne Cup in Figure 2, the mode is 5 years. We could say that 5 years is the typical age, but looking closely at the plot we see that the frequency for 5 years is only one greater than the frequency for 4 years. Hence we might want to hedge our bets and say the typical age is '4 or 5 years'!

There are also other ways to talk about typical values in data sets. For example we

[2] Colt: a male horse under the age of four. Filly: a female horse under the age of four. Mare: a female horse four years old and older. Stallion: a non-castrated male horse four years old and older. Gelding: a castrated male horse of any age.

could talk about putting numerical data in order from the smallest to the largest and finding the value in the middle. This value is called the median. That would give us a feeling for what is a typical value in the middle of the set. If we looked for the median of the ages of the winners of the Melbourne Cup, with 153 values, we would want the value of the 77th Age, the age with 76 Ages on either side. Looking at the plot in Figure 2, we see that there are 66 values in the bins for 3 and 4 years of age. That means that the 77th value is in the 5–year-old bin. So the middle value in the data set, the median, is 5 years.

Returning to Figure 1, is it possible to talk about the median sex of the winners of the Melbourne Cup? One might claim that Colt is the median because it is in the middle. The 77th value from the left in the plot, however, lies in the Gelding block. Being the centre category in a bar chart does not make a value the median. Suppose the Sex categories were replotted alphabetically as in Figure 3. Then the 77th value would lie in the Mare block. What would this tell us? Only that it does not make any sense to talk about the median of a categorical data set. Because the data cannot be ordered numerically, there can be no middle value. Leavy, Friel, and Mamer (2009) give other examples of students trying to find the median of categorical data sets.

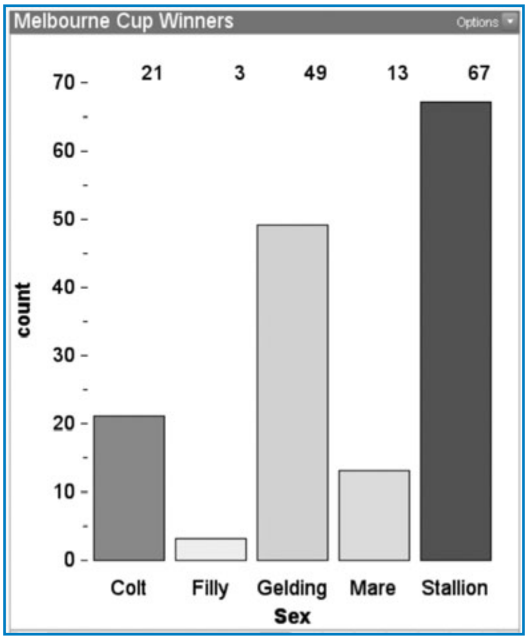

Figure 3. Sex of Melbourne Cup winners (ordered alphabetically).

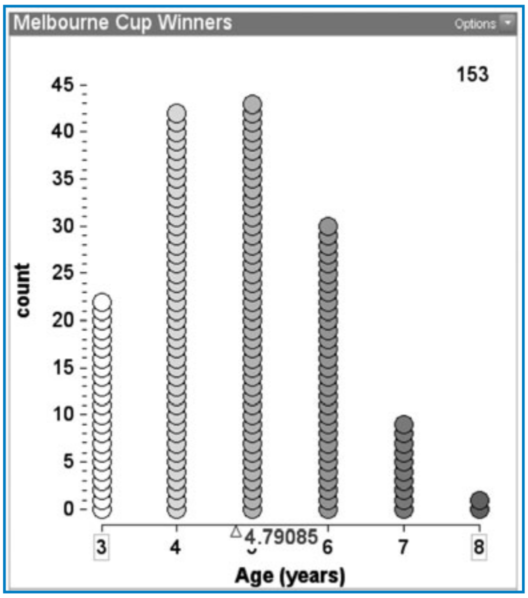

Figure 4. Ages of Melbourne Cup winners with the mean marked on the axis.

As numerical data become more detailed in their representation and clustered or spread out, the most complex way of considering what is typical is likely to be the calculation of the mean. The mean is the total of all of the numerical values in the data set divided by the number of numerical values. Physically this would be the balance point if all of the values could be placed on a scaled ruler with a fulcrum at this numerical value, the mean. Basic activities based on the mean as balance point are found in O'Dell (2012) and Hudson (2012).

The mean Age of the winners of the Melbourne Cup turns out to be 4.79 (733 divided by 153). Notice that this number, shown in Figure 4, is not one of the ages of the winners. This might be considered annoying if we wanted to use the mean to talk about the typical age of a winner, from the data given, which are in whole numbers. So perhaps the best way to talk about the typical age of a winner of the Melbourne Cup would be to use the median, if we wanted a single value that would be the most intuitive to the general people interested in the Cup.

When data sets become more detailed numerically, often the mean is the best way to describe the data in one number because it takes into account every number in the data set. Consider the plot in Figure 5 of the Weight carried by the winners of the Melbourne Cup. It is quite crowded in the middle and we see that the mean is marked as 51.2 kg. It is pretty easy to imagine this number as a balance point because the values are fairly evenly distributed on either side of it. It also might seem reasonable in this case that 51.2 is one of the values in the data set.

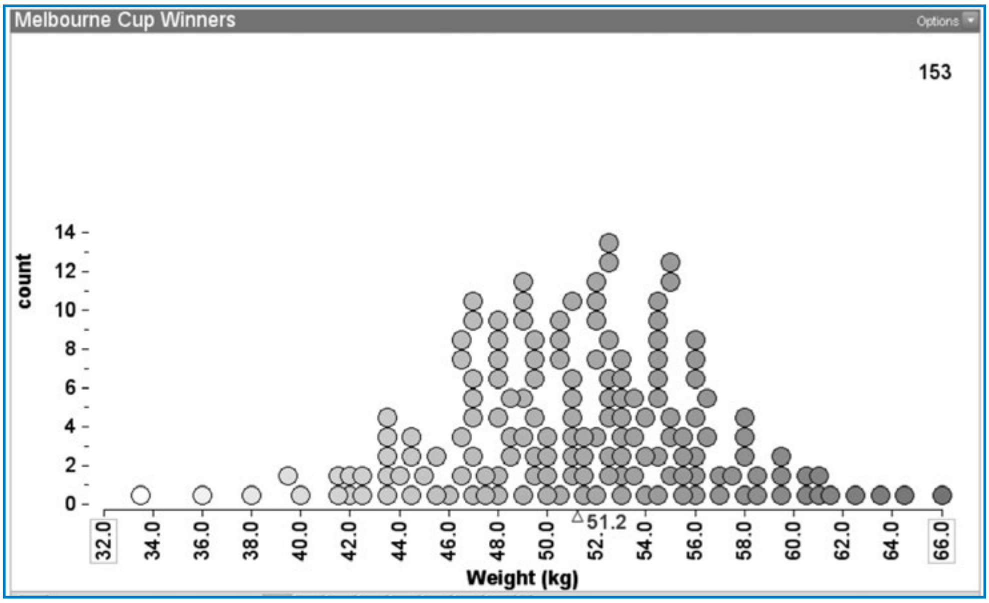

Figure 5. Weight carried by the Melbourne Cup winners.

Would the mode be useful to describe the typicality of this data set? We would need to separate the data better to tell. Looking at the plot in Figure 6 we see that there are three modes (52.5, 53.0, and 54.5). These values, however, have only one greater frequency than four other values (47.0, 48.0, 51.0, and 56.0). The mode is hence not very useful in telling us about the typicality of the Weights carried by the winners of the Melbourne Cup, except that it does tell us that there are quite a few values that are repeated about the same number of times in the centre of the data set.

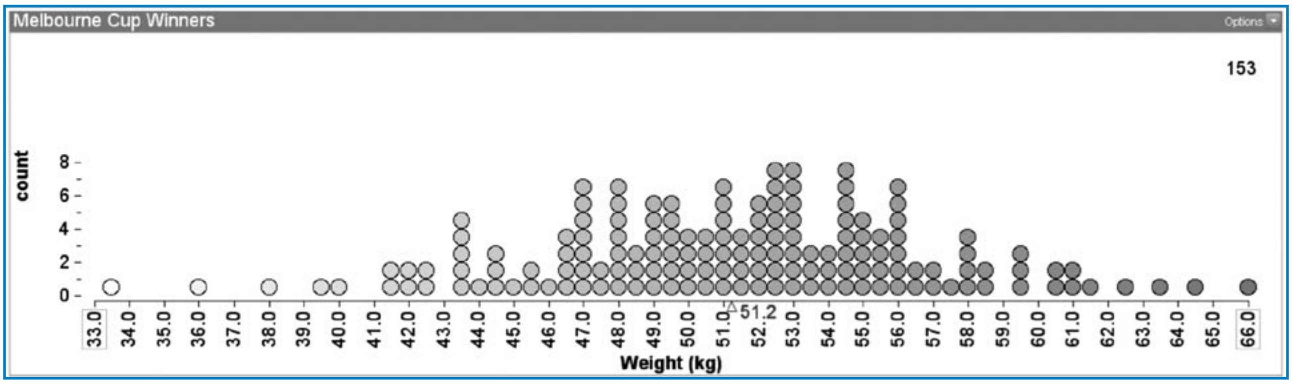

Figure 6. Weight carried by Melbourne Cup winners showing 3 modes.

What would the median tell about the typical value/s in this data set? As we can see in the plot in Figure 7, the middle value is one of the 51.5 values, which is very close to the mean value of 51.2 kg. Now we see that again (as for Age) the mean is not one of the values in the set.

The closeness of the mean and the median tells us that the distribution of the Weights is relatively even on both sides of these middle points. In other words the distribution is fairly symmetric. Looking at the plot for the Winning Margin (measured in 'lengths'), however, we see a different shape. Because there have been many close finishes to the Melbourne Cup, in Figure 8 there are many values close to zero.

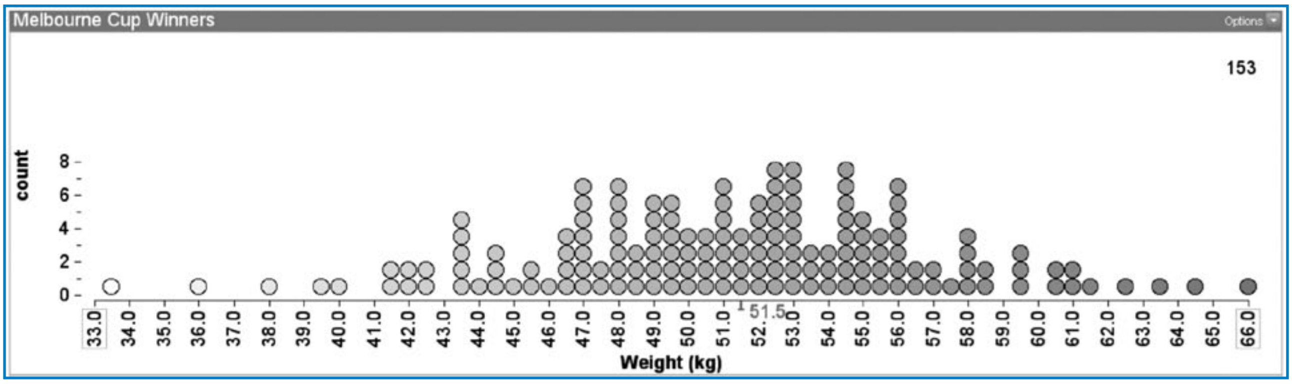

Figure 7. Weight carried by Melbourne Cup winners showing the median of 51.5 kg.

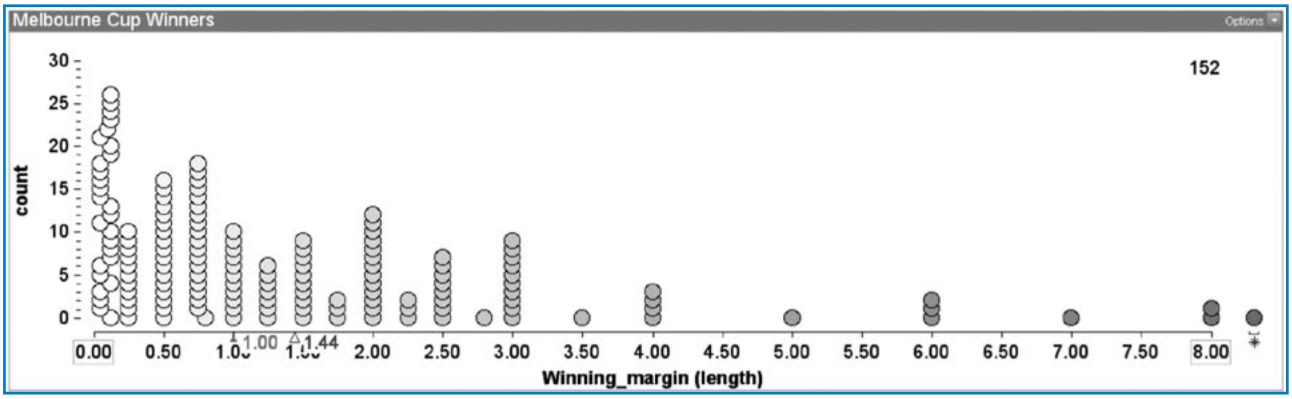

Figure 8. Winning margin, measured in lengths for the Melbourne Cup.

For Winning Margin we see that the median is 1.00 and the mean is 1.44. This is because the median takes into account the order of the data but not their exact values. Here the median is halfway between the 76th and 77th values (there is one missing value); but both values are 1.00, hence this is the median. The mean is larger because it takes into account every value in the data set and is influenced by the large values on the right. The median is not influenced by the actual values, only that they are to the left or right of the 76th and 77th values. The mode is of little use here because of the many values very close to 0. All we can say with reference to this is that it is typical for races to have a close finish! This comes from looking at the entire distribution.

Returning to Figure 1 showing the Sex of winners of the Melbourne Cup, we realise that it would be impossible to try and calculate a mean. One cannot multiply Filly × 3, Mare × 13, etc., add them together, and then divide by 153! The mean is not used at all for categorical data sets, nor is the median.

It is more likely that the categorical variables are used to separate the data set in ways that are of interest when exploring the numerical variables. For example, consider the plot in Figure 9 where the Weight carried by the winner is on the horizontal axis and the Sex of the winner is on the vertical axis, creating 5 sub-groups of data. In this case we notice that we can order the Sex of winners to show an increasing mean value for Weight carried from Filly, to Colt, to Mare, to Gelding, to Stallion. An expert could probably explain why this might be so! Categorical variables have a very important part to play in statistics but it is not related to being able to produce a mean or median value as a representative statistic.

Some numbers are used as values for a variable but are treated as categorical rather than numerical. Consider for example the variable Year—for the year that the Melbourne Cup was run—containing the values from 1861 to 2013. We could work out the median year or calculate the mean year by adding the years and dividing by 153. In this case, the median would be the middle year, 1937, which is meaningless in the context of the Melbourne

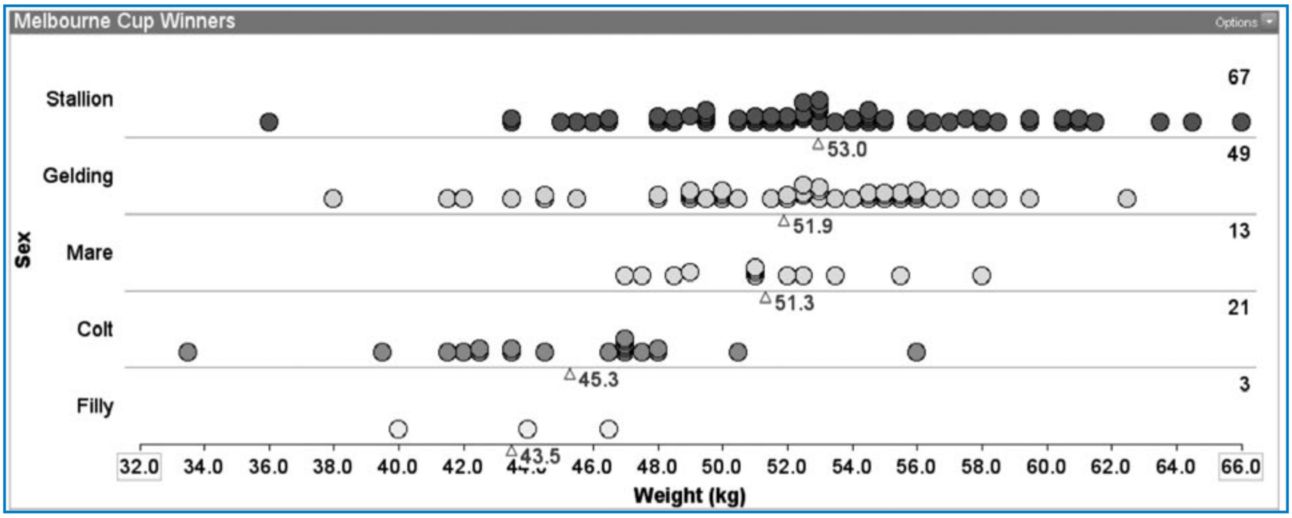

Figure 9. Weight carried by Melbourne Cup winners separated by Sex.

Cup. Because the data are totally symmetric the mean will have the same value of 1937. Further, the mode for the variable Year is similarly useless because every year in the data set has a frequency of one! It tells us nothing about typicality.

A categorical variable like Year is useful to us precisely because it is uniform along a scale and it helps us keep track of other variables in a 'time series'. For example, if we look at the Weights carried by the winners of the Melbourne Cup, we see what seems to be a reduction in the variation in weights over the years. Hence although Year is not a variable of much interest on its own (except to keep track of how many years the Cup has been run), it is useful in other ways.

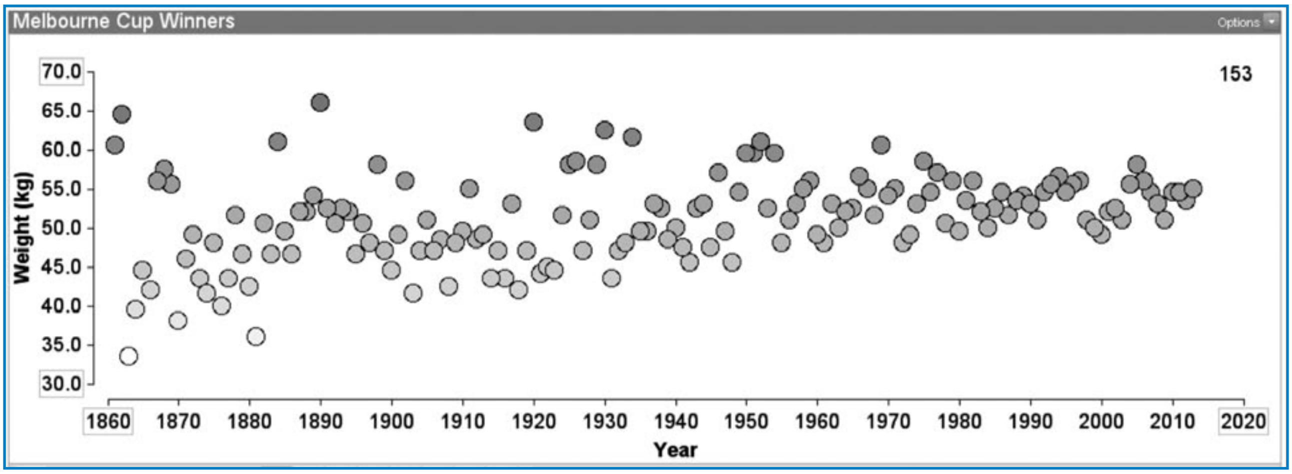

Figure 10. Plot of Weight carried each year by the Melbourne Cup winner.

It is interesting to note the increasing complexity of the procedures associated with the concepts of mode, median and mean, in terms of their use of the information contained in a data set, although each potentially produces only one number.

- The mode only looks for 'most' frequent values, so could be identified visually from an appropriate plot in one step.
- The median looks for 'middle' and uses all of the values in the data set but only their order. Only counting is needed once the values are placed in order. There are therefore two steps: place the values in order and then count to find the middle.
- The mean looks for 'balance' and uses all of the values in the data set including their numerical values. They need to be added carefully and

then divided by the number of values. Even with a calculator this can be a tedious job. There are three steps: record values, add values, divide the total by the number of values.

A wider view of 'typical'

For numerical data we can also take a somewhat broader view of typical by considering the 'middle' values of data, rather than just looking at a single value. The box plot uses the median as its centre because it is the value half-way through the ordered data. A rectangular box is then created to represent the quarter of data on either side of the median. The box hence 'covers' the middle half of the data, which gives us a general appreciation of the typical values in the centre of the distribution. At either end of the box a whisker is drawn to the extreme values at the ends of the data set, marking the lowest quarter and the highest quarter of the data. This gives us a feel for the overall variation in the data within which we are focussing on typical values.

Box plots vary greatly depending on the shape of a data distribution. This is seen in Figure 11 with the box plots for the Weight carried by winners of the Melbourne Cup and the Winning Margin in the race. On one hand, the box representing the middle half of the data is in the middle of the plot for Weight, and the median is in the middle of the box, showing us that the data for this variable are quite symmetric (check this in Figure 5). On the other hand, the middle half of the data for Winning Margin is nowhere near the middle of the plot, indicating a data set that is skewed away from zero. The plot shows the lower 25% of the data squeezed up near zero and the upper 25% of the data very spread out from 2.0 to 8.0 lengths. More description of box plots is found in Watson (2012).

Box plots are useful for a quick summary comparing the shape of data when separated by a categorical variable. Figure 12 shows the box plots for Weight carried for each of the categories of the Sex variable. The impression is slightly different from the representation in Figure 9 because the

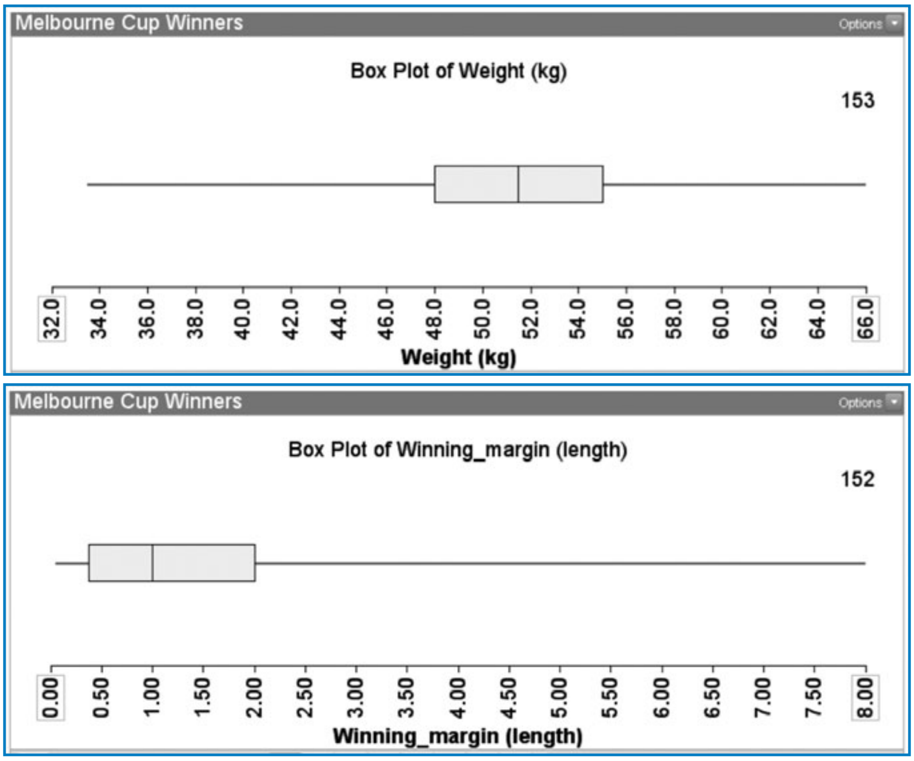

Figure 11. Box plots for Weight carried and Winning Margin for Melbourne Cup winners.

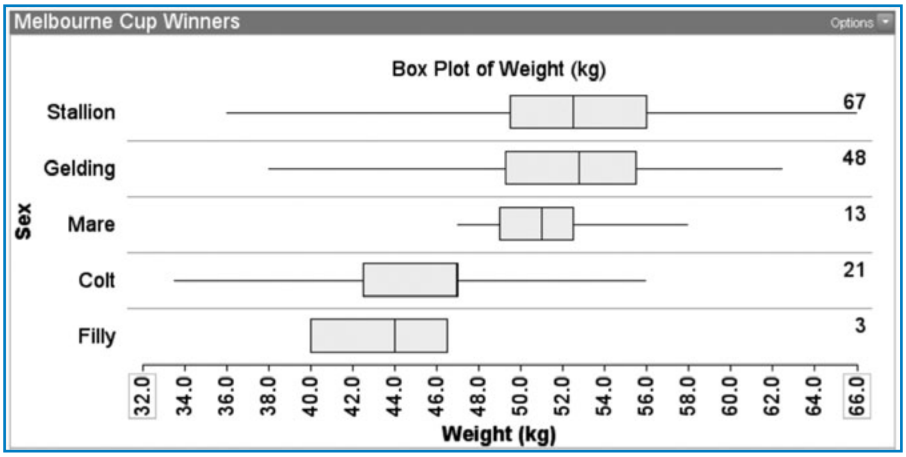

Figure 12. Box plots for Weight carried by Melbourne Cup winners separated by Sex.

middle of the box is the median in Figure 12, not mean as in Figure 9. Would this make us change our minds about whether Stallions carried more weight than Geldings when winning the Melbourne Cup? The variation in the data is greater for Stallions than Geldings and we see the difference in the impression of typical when the mean is used rather than the median.

Starting with the idea of 'typical' as the mode, we have progressed through considering the median and mean, to expanding 'typical' to include the middle half of a numerical data set. Although the box plot is not introduced in the *Australian Curriculum: Mathematics* (ACARA, 2013) until Year 10, students should become cognizant of the critical importance of the overall shape of a data set when considering what is typical. What is typical for the Weight carried by Melbourne Cup winners is different from what is typical for their Winning margins. The Weight is typically symmetric whereas the Winning Margin is typically skewed.

Whether one is 'into' horse racing or not, every year the Melbourne Cup provides a genuine Australian context within which it is possible to explore the important concept of typicality in relation to three measures of centre, two types of data sets, and the shapes of graphical representations for the data sets. By the time students emerge from middle school, they should have built intuitions about what is typical in data sets that go far beyond a procedural definition of mean or median found in a text book.[3]

References

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2013). *The Australian curriculum: Mathematics, Version 4.1, 1 February 2013*. Sydney, NSW: ACARA.

Hudson, R. A. (2012). Finding the balance at the elusive mean. *Mathematics Teaching in the Middle School, 18*, 301–306.

Jacobbe, T. & Fernandes de Carvalho, C. (2011). Teachers understanding of average. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds), *Teaching statistics in school mathematics: Challenges for teaching and teacher education* (pp. 199–209). New York: Springer.

Konold, C. & Miller, C. D. (2011). *TinkerPlots: Dynamic data exploration* [computer software, Version 2.0]. Emeryville, CA: Key Curriculum Press.

Leavy, A. M., Friel, S. M., & Mamer, J. D. (2009). It's a fird! Can you compute a median of categorical data? *Mathematics Teaching in the Middle School, 14*, 244–351.

O'Dell, R. S. (2012). The mean as balance. *Mathematics Teaching in the Middle School, 18*, 148–155.

Watson, J. (2012). Box plots in the Australian curriculum. *The Australian Mathematics Teacher, 63*(3), 3–11.

Watson, J., Beswick, K., Brown, N., Callingham, R., Muir, T. & Wright, S. (2011). *Digging into Australian data with TinkerPlots*. Melbourne: Objective Learning Materials.

[3] Assessment, including NAPLAN testing, should reflect this understanding!