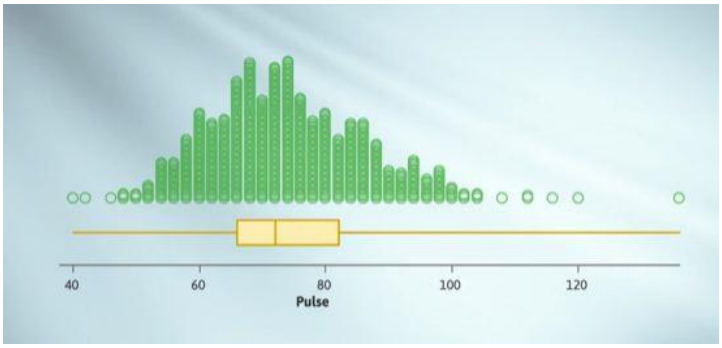
THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

WEEK 2

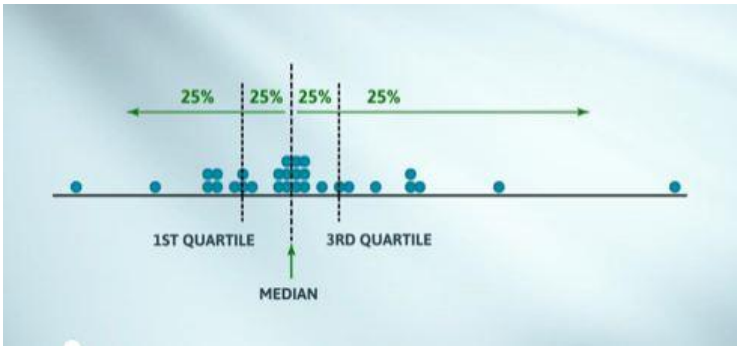
2.9 FEATURE SPOTTING by Chris Wild

Hello again. We're continuing our saga about numeric variables. Last time, I told you that the main things we look for in dot plots are centre, sprad, shape, and oddities.

In this video, I'll talk about the last three. I'm assuming that you've already read the article, Features of Numeric Variables.

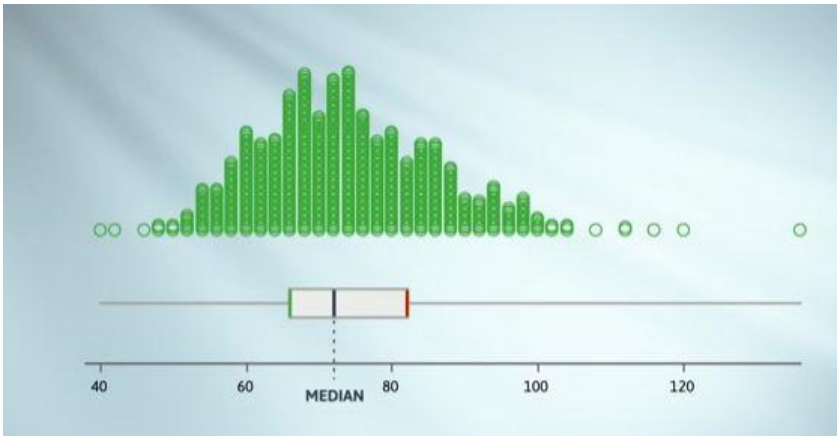


If you haven't seen them before, you may be wondering what these boxes are at the base of our iNZight graphs. I'll explain them now.

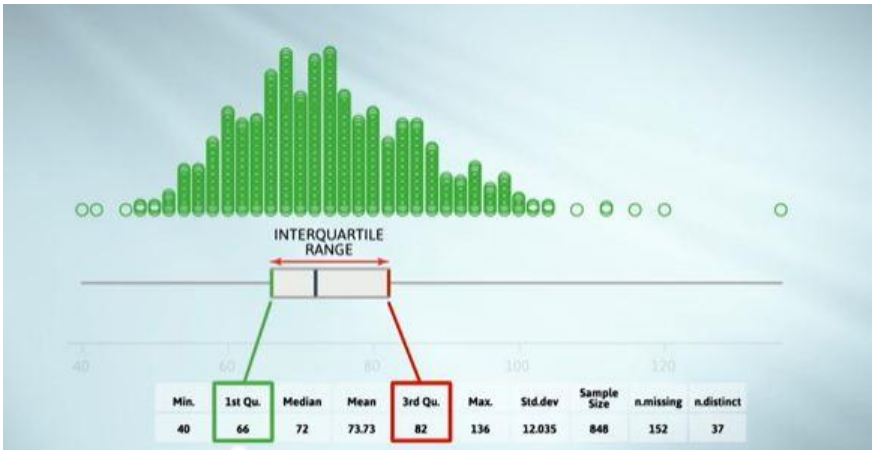


When we cut the data in half (half above, half below) this gives us the median. The position that divides the bottom half in half -- so that a quarter of the observations lie below, and three-quarters above it -- is the first (or lower) quartile.

Similarly, the third (or upper) quartile divides the upper half of the data in half. So that three-quarters lies below, and a quarter of the observations lie above. Circling this gives us the middle half. Putting it together gives us the box plot.

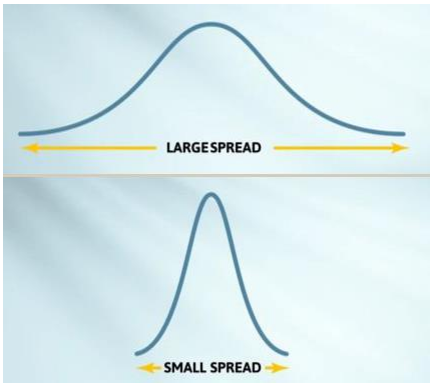


iNZight routinely draws the box plot under the dot plot, so that we can see where these values are. We can see here that the median pulse rate is a little over 70. And the middle 50% of our people had pulse rates somewhere between about 65 and 82.

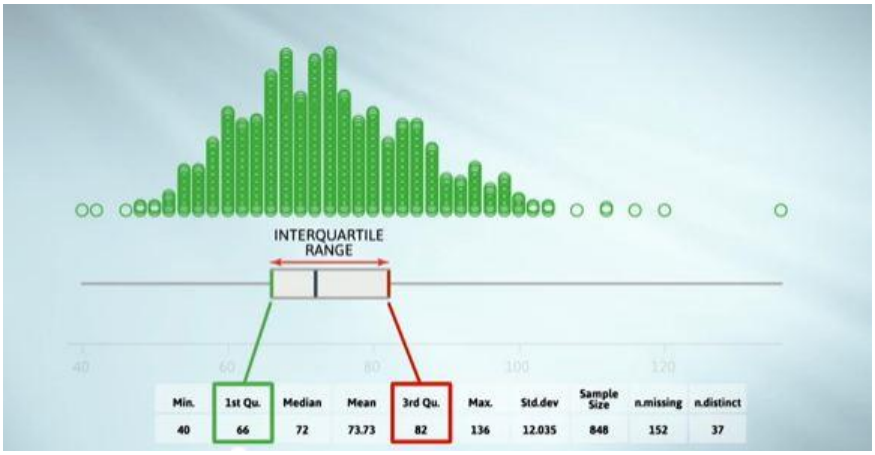


Here are the actual values. The length of the box, which is the difference between the third and first quartiles, is called the inter-quartile range. The inter-quartile range is one of the standard measures of spread.

Here is the basic idea of spread.



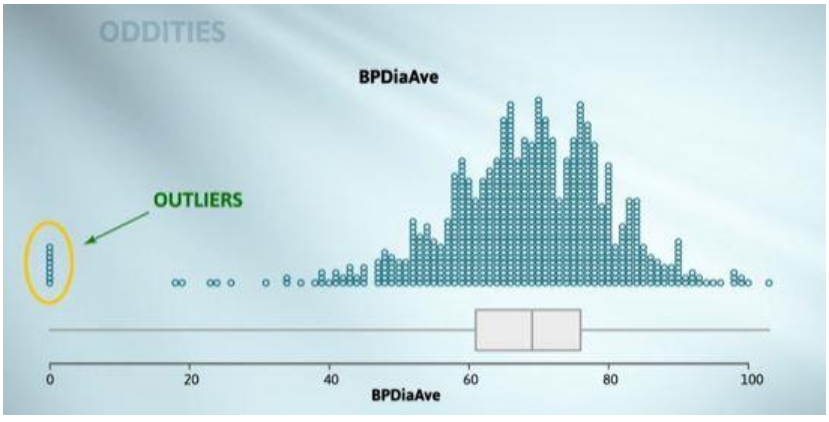
The concept of spread answers the questions, "How spread out are the observations along the scale?" or equivalently, "How much did these observations vary?"



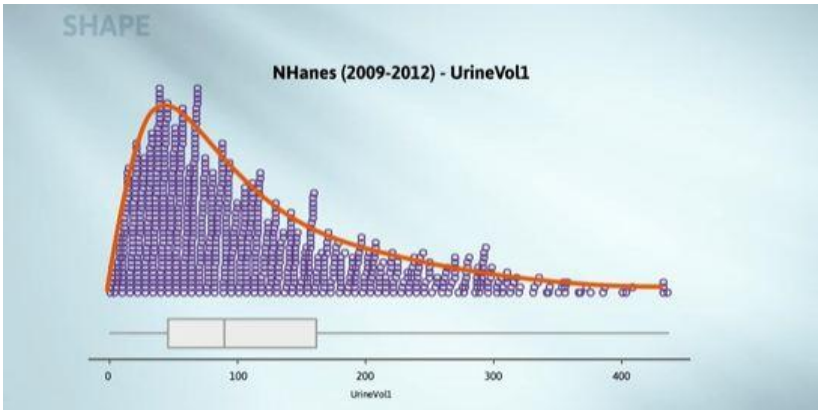
Looking back at this plot, our pulse rates are centred at about 72. They're quite variable, ranging all the way from about 40 to nearly 140. With the vast majority falling between about 45 and 105. The shape is a fairly symmetrical mountain shape.

Last in our list of features is oddities.

By oddities, we just mean anything that looks strange or odd. While we've come to this last, these are often the first things that pop out at us when we look at a graph, as with this set of 0 readings for diastolic blood pressure.

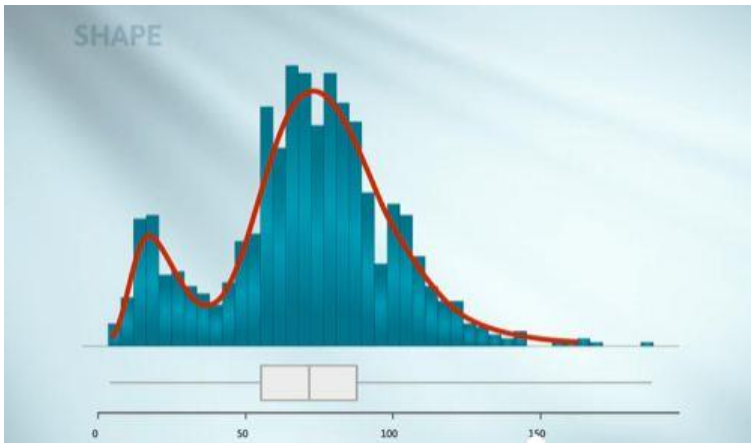


Outliers are data points that are sufficiently far from the general pattern that they look suspect-- enough to make us worry, are these values real or are they mistakes?



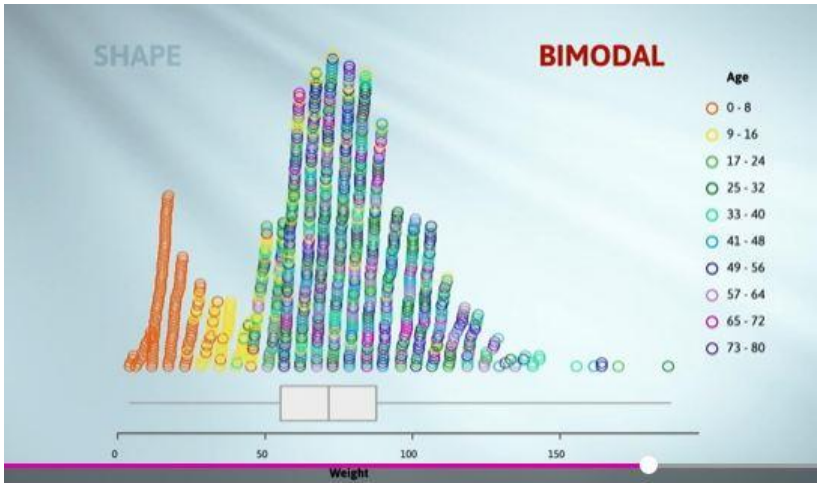
We're now focusing on interpreting shapes. Whereas pulse rates form a symmetric mountain shape, the results for the first urine volume test are severely skewed-- bunched up towards 0 on the left-hand side, with a long, stretched-out upper right-hand tail. We call this shape, positively skewed.

Something that'll probably look strange to you in this dot plot is the way the towers of points lean. It's not something to worry about. It's a consequence of iNZight stacking points that are very close together, but not identical.



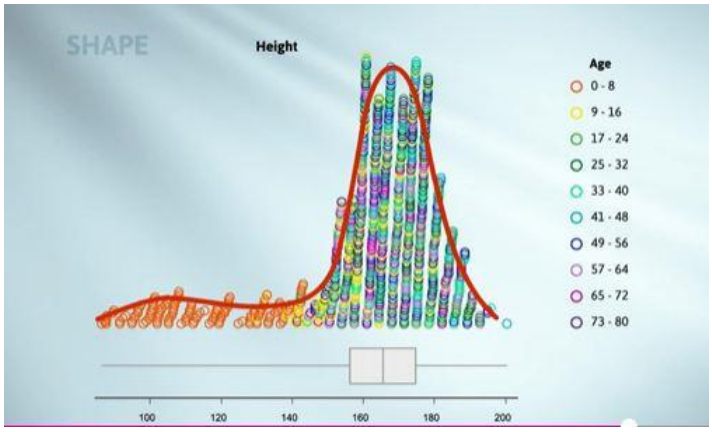
Weight

Here with weight, we see two mountain peaks. A larger one centred at about 75, and a smaller peak to the left-hand side, centred at somewhat over 10. If you can see two mountain peaks, the shape is said to be bimodal. When we see this, we suspect that what we're seeing is a mixture of at least two different groups, and start wondering what they are. Any ideas?



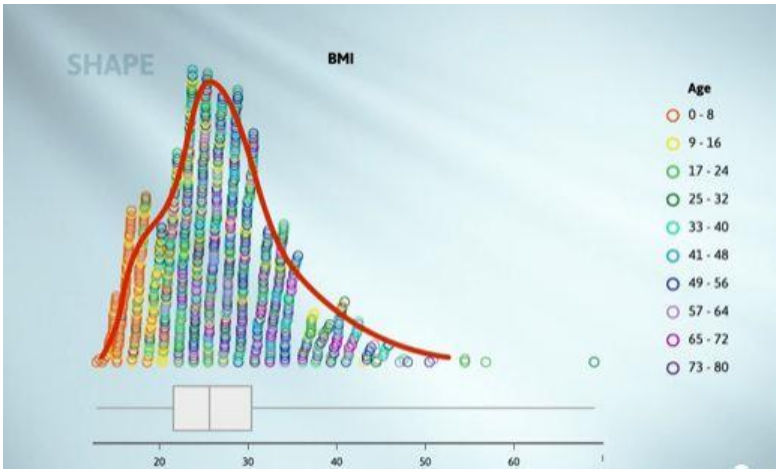
Weight

Here, I've coloured the individual points by their age. See how the colours relate to age? The younger ages are the oranges and yellows. And the older ages are the blues and purples. The lower mountain has the orange, then yellow, shades. It's made up of the children and teenagers.



Height

Interestingly, we don't get a second mountain, to anywhere near the same extent, with heights. The left tail is much flatter, here. I wonder why that is. The left tail has the children, in orange. And the teenagers, in yellow, are much more mixed in with the adults, here.



Body mass index (BMI)

Body mass index is a measure of how heavy you are for your height. It's used as an inexpensive measure of over- or underweight in health studies. We can see an age gradient on the left-hand side. The younger ones are lighter for their heights. But it still doesn't separate out into two peaks, like the weights did. It's also somewhat right-skewed, but less extremely than urine volume was.

All of these things are useful for grasping how things are for a single variable. They're even more useful for getting a handle on how things change. As in this plot, constructed to investigate the relationship between first marijuana use and educational attainment level. But we're getting ahead of ourselves, here.

Finally, I'll leave you with these questions, to remind you of the ideas we've just been covering.

- What sort of "oddities" should we look for and what sort of questions should we ask when we see them?
- What is the defining property of a median? the first quartile? and third quartile?
- What information does the box plot add to the dot plot?
- What is the interquartile range and how does it relate to the box plot?
- What is an "outlier"?
- What does "positively skewed" mean?
- What does "bimodal" mean?