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

WEEK 2

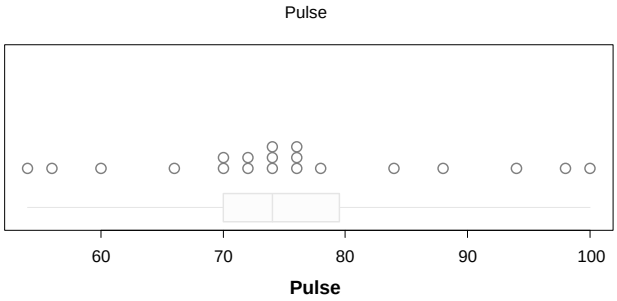
2.7 NUMERIC VARIABLES by Chris Wild

Earlier in the week, we looked at plots and summaries of a categorical variable. All we could do is count the numbers of people that fall into each category, and display them in various ways.

Our primary plot was a bar chart of the percentages falling into each category. This time, we'll look at a single numeric variable. They're much richer than categorical variables, in terms of what we can look for, and what we can see. We plot them quite differently.

Height	BMI	BMI_WHO	Pulse	BPSysAve	BPDiaAve
117.5	16.7	12.0_18.5	NA	NA	NA
167.7	23.1	18.5_to_24.9	72	86	60
188.6	28.3	25.0_to_29.9	94	125	72
151.3	23.1	18.5_to_24.9	98	117	76
175.6	29.8	25.0_to_29.9	76	113	47
168.5	23.2	18.5_to_24.9	74	120	83
181.2	30.4	30.0_plus	60	134	92
174.7	26.1	25.0_to_29.9	74	100	76
118.8	15.2	12.0_18.5	NA	NA	NA
171.5	26.2	25.0_to_29.9	70	130	83
170.5	27.7	25.0_to_29.9	56	124	71
153	18.4	12.0_18.5	66	87	55
182.1	25.6	25.0_to_29.9	88	138	88
167.5	23.8	18.5_to_24.9	74	116	64
...	...	...	...	...	...

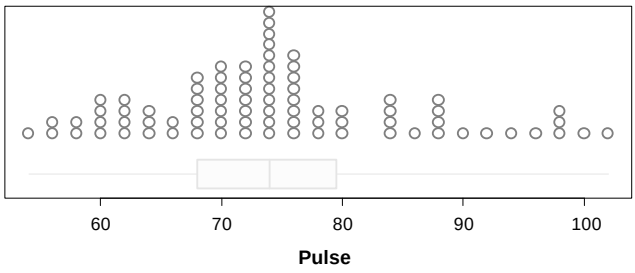
We'll begin with the resting pulse rates of people in the NHANES data set. What's your typical resting pulse rate? Mine's about 70. Very fit athletes have pulse rates as low as 40. Here are the first 20 pulse rates in our NHANES data set.



Stacked dot plot showing the first 20 values

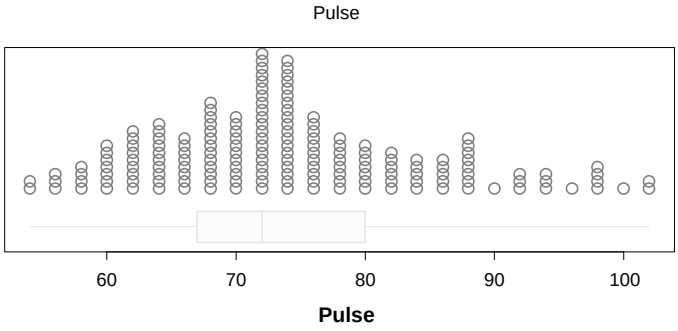
This is the simplest plot you can construct of a set of numbers. It's called a stacked dot plot. Each value is plotted against the scale using a dot-- so each dot represents a person. This dot represents someone with a pulse rate of 54. Someone with rate of 56. And so on.

When people have the same value, we stack them one above the other, so we can see how many there are. Hence, the "stacked" in stacked dot plot. Here we can see two people with a pulse rate of 70. And three with 74. Ignore the faded box shape at the bottom, for now.



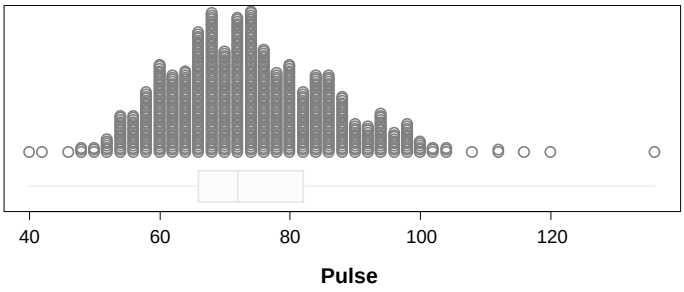
Stacked dot plot showing 100 values

These pulse rates go all the way from 54 to 100. Here we've plotted the first 100 values. We can see where we're getting lots of values, because the stacks have piled up high there. Similarly, we can see values where no one, or only a few people have.



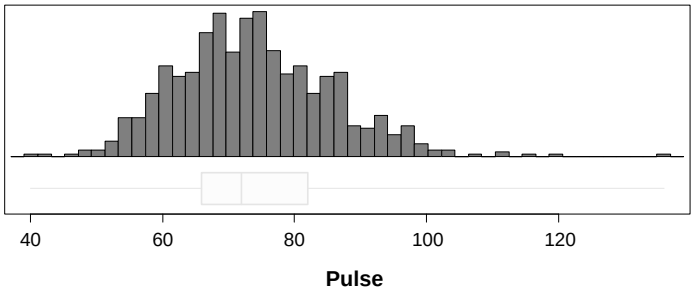
(Stacked) dot-plot showing 200 values

At 200 data points, with the plotting window being used here, there isn't room to stack them one on top of another, because we'd run out of vertical space. Instead, we overlap them just enough so that the highest stack still fits in the available space.



(Stacked) dot-plot showing 1000 values

By the time we get to the first 1,000 data points, the only individual dots that we can see are those at the extremes, where very few instances occur.



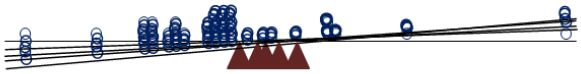
A *histogram* of the same points

iNZight defaults from drawing dot plots to drawing histograms, like this one, at about 2,000 observations. Histograms are faster to draw for large data sets. But you'll see they're giving us the same shapes to interpret.

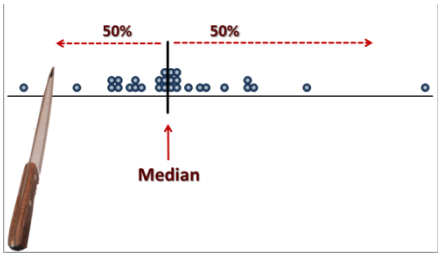
But the programme lets us override the default, and ask for the easier-to-understand dot plots. Here the main things we look for in dot plots-- centre, spread, shape, and oddities.

We'll start by looking at centre.

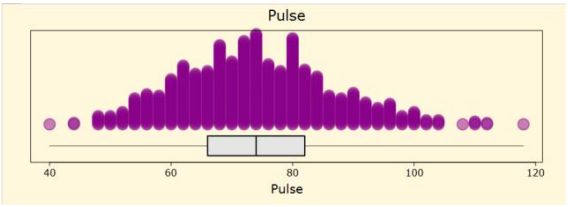
The Mean is where the plot balances



One way to think about the idea of centre is the point of balance-- where the dot plot balances. It turns out that the balance point is the ordinary, everyday average-- what statisticians call, the mean. If you're looking at a graph, and wondering where the mean is, you should ask yourself-- where would this balance?



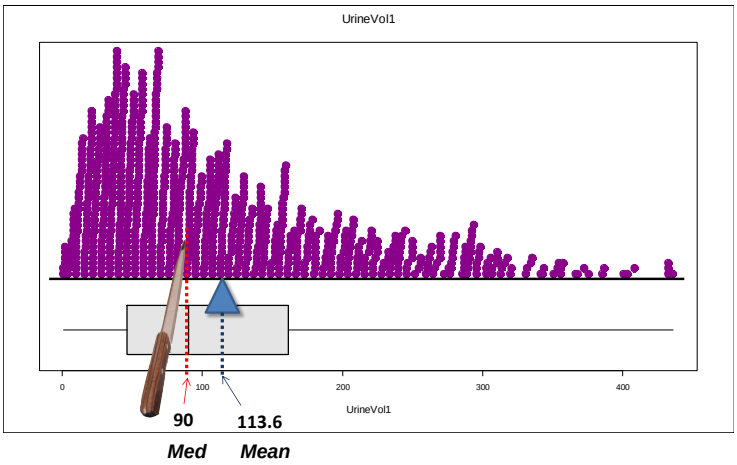
The other common notion of centre is the median. The median divides the data in half-- with half of the observations above it, and half below it.



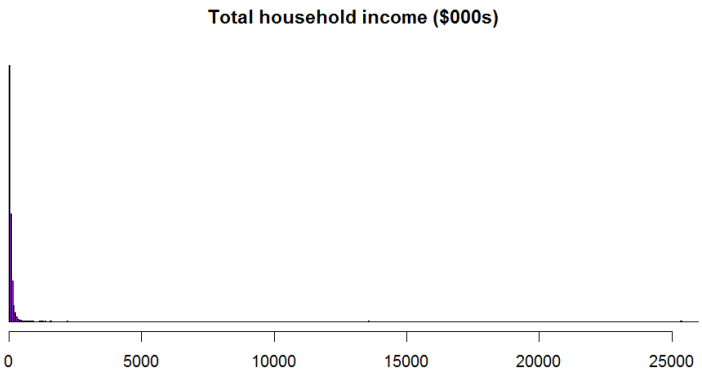
The graph of pulse, here, looks like an approximately symmetric mountain. When the data is nearly symmetric, the mean and the median are almost the same. The median here is 72, and the mean 73.7.

If the data is perfectly symmetrical about a central axis, we have a very well-defined notion of centre, and all measures of centre-- including the mean and the median-- give the same answer.

For the NHANES variable, UrineVol1, the median and the mean are quite different--90 versus 115.6.

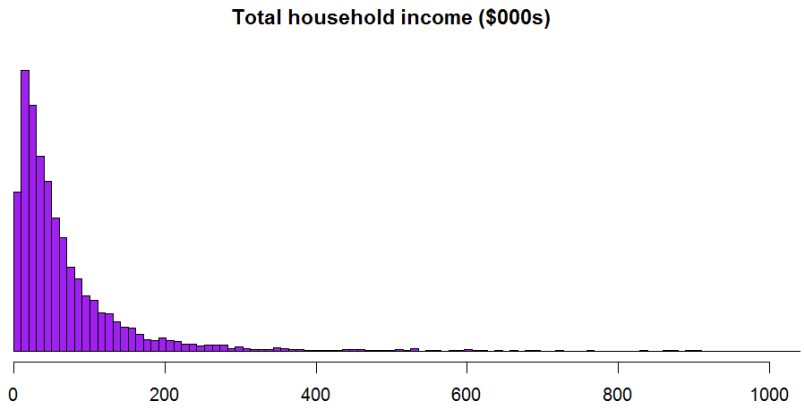


This plot is far from being symmetric. This sort of shape is said to be skewed. With skewed data, there's no sharp notion of centre. Different ways of thinking about the idea of centre produce rather different answers, as we see with the mean and the median, here.

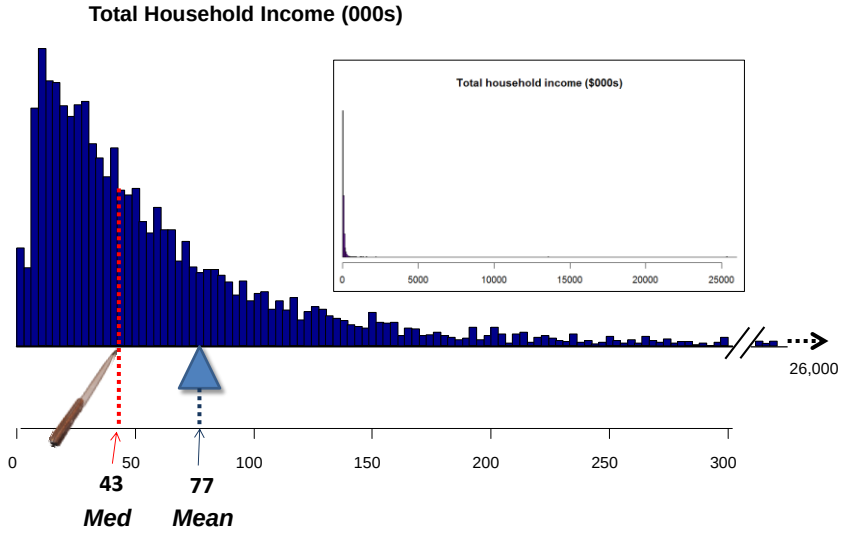


This plot shows an extreme case. The shape here is enormously skewed, squashed into what almost looks like a single spike near 0. It is showing total household incomes for the 2006 US Health and Retirement Survey. Household incomes go all the way up to $26 million.

Next, we'll look more closely at the data squashed into the left-hand side of this plot.



We've zoomed in on just the incomes up to $1 million. And we can see that nearly all household incomes are under about $300,000.



Here we've zoomed in further-- to just the incomes up to $300,000. We're showing the median and the mean for the whole data set. There's a big difference between the two. The median is $43,000. And the mean-- which is where the dot plot balances-- is nearly twice as big, at $77,000. This mean doesn't represent the common experience. Only 27% of these incomes fall above $77,000, with 73% below them.

So which measure of centre should we use? That depends on what we want to use it for. For incomes, and many highly-skewed variables, the median tends to be preferable because it better represents the common experience, or what is typical. Half of the people get less. Half get more.

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

QUESTIONS

- How is a stacked dot plot constructed?
- What are the four types of things we look for in dot plots?
- What is the common name for the mean?
- How does the mean relate to the dot plot?
- When do the mean and median tend to be very similar, and when can they be quite different?
- Why might we prefer the median to the mean for personal incomes?