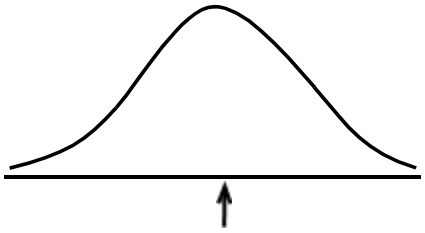
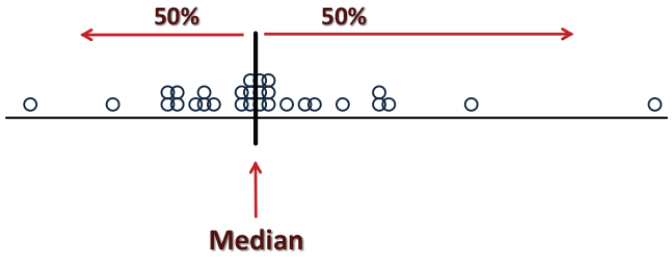
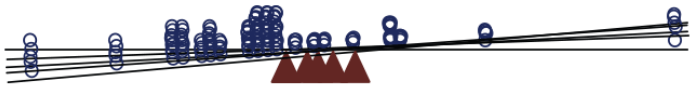
THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

DATA to insight
An Introduction to Data Analysis

Features of numeric variables

- Centre
- Spread
- Shape
- Oddities

In what follows we use summary measures on the pulse rate variable in the NHANES 2011-12 for illustration.

Centre	Measures	Notes
Where the **"middle"** of the set of observations is.  The two most commonly used measures of centre are the **median** and the **mean**.	**Median** Cuts in half. Half of the observations are above the line and half are below.  **Mean** This is statistical name for the ordinary, everyday average. It is where the dot plot balances. 	Summary of Pulse: Min. 40 · 1st Qu. 66 · Median 74 · Mean 73.87 · 3rd Qu. 82 ▪ If we had to summarise all the observations on a variable as a single number then we'd want to use a measure of their "centre". ▪ If the shape is roughly symmetric the mean and the median will be approximately the same. ▪ If the shape is strongly skewed there is no single compelling notion of "centre" and the mean and median can be quite different. ▪ Means can be quite badly affected by outliers in smaller data sets whereas medians are not.

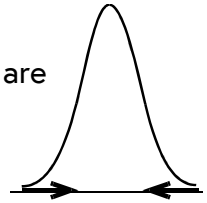
Spread

How **"spread out"** the observations are along the scale.

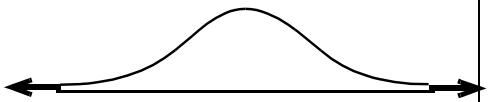
Equivalently, how **variable** the observations are or how **different** they are from one another.

Smaller spread

The observations are **less variable** i.e., **less different** from one another



Larger spread



The observations are **more variable**, i.e. **more different** from one another.

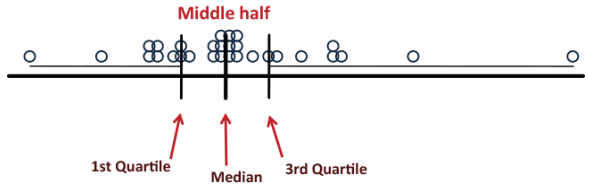
The two most commonly used measures are: the **interquartile range** (IQR) & the **standard deviation**.

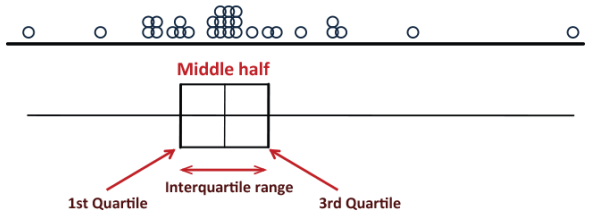
Measures

Summary of Pulse:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.dev	Sample.Size	n.missing	n.distinct
40	66	74	73.87	82	118	12.283	853	147	35

Interquartile Range (IQR). To understand the IQR we need first to understand the **quartiles**.





Recall that the median is "the value that cuts the data in half" (in the sense that half of the observations fall below it and half above). In the same way, the 1st quartile, the 2nd quartile (=median) and the 3rd quartile cut the data up in quarters. Cutting the bottom half of the data in half to gives the **1st** (or **lower**) **quartile** and cutting the top half of the data in half gives the **3rd** (or **upper**) **quartile**.

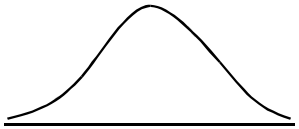
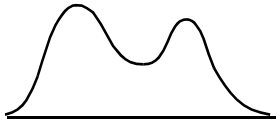
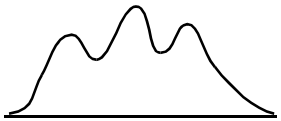
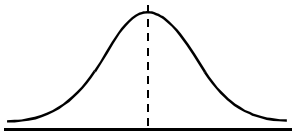
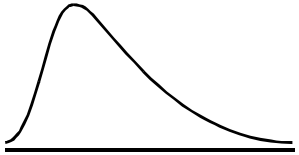
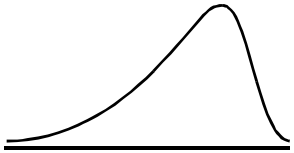
The **interquartile range** is the range of values spanned by the middle half of the data
(IQR = 3rd quartile – 1st quartile.)
It is also the **length of the box in the box plot**.

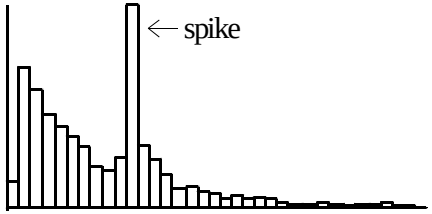
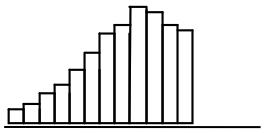
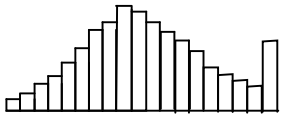
Standard deviation

You won't go too far wrong in thinking of it as **the average of the distances between the points and the mean**.

When the dot plot is a roughly-symmetric mountain shape then typically **about 66% (two thirds) of the data points fall within one standard deviation either side of the mean**.

In detail, the standard deviation is **the square root of the average of the squared distances between the points and their mean**. Average the squared distances and then take the square root of the answer.

Shape	Notes
Modality  **uni**modal = **one** "mountain"  **bi**modal (**2** "mountains")  **tri**modal (**3** "mountains") etc. **Symmetry and skewness**  Symmetric (& unimodal)  Positively skewed  Negatively skewed	▪ We do not pay much attention to details of shape unless the departures from "roughly symmetric and unimodal" are fairly extreme or they are being seen clearly in very large data sets. ▪ Having more than one mode suggests the existence of distinct groups (an investigator would try to find defining characteristic). ▪ Recall that if the shape is strongly skewed the mean and median can be quite different. ▪ **Positive skewness** is sometimes called **right skewness**, and **negative skewness** called **left skewness**.

Oddities - Should prompt, "*What's going on here?*" *(indicators of possible problems with the data or opportunities for discovery)*			
Name	**Symptoms**	**Usual Suspects**	**Follow-up**
Outlier(s)	Data points sufficiently far from the general pattern that they look suspect. 	▪ Mistake/Error. ▪ Something real and unexpected.	▪ First try to resolve whether it is real. Go back to original sources and correct it if possible. ▪ Try to find real-world cause (may lead to a discovery).
Gaps and Clusters		▪ Existence of distinct groups.	▪ Try to find defining characteristic.
Spike		▪ Mistake/Error. ▪ Nearby values have sometimes been rounded to this (even more often the cause when there are several spikes). ▪ Something real and unexpected.	▪ Try to resolve whether it is real. Go back to original sources and correct it if possible. ▪ Is this a problem for intended analysis? ▪ Try to find real-world cause (may lead to a discovery).
Truncation	 Looks like the end has been chopped off.	▪ Everything with larger values has been eliminated.	▪ Why? Is this a problem for intended analysis?
Truncation with spike		▪ All larger values have been set to spike value.	▪ Why? Is this a problem for intended analysis?