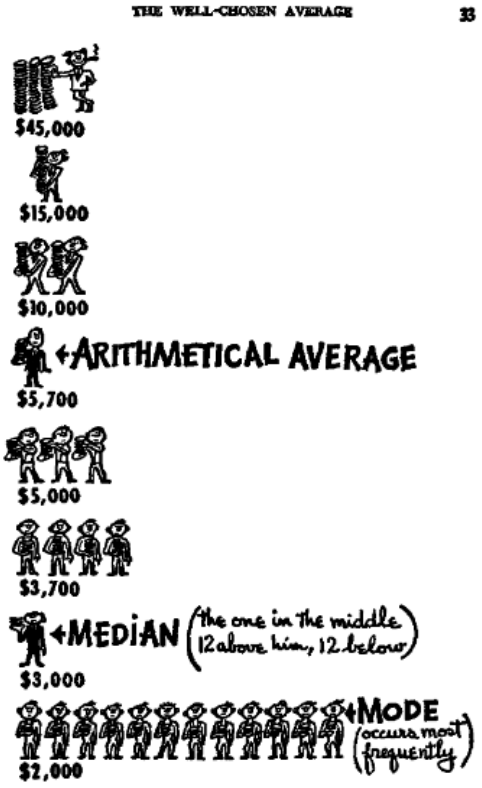
NCEA Level 2 Mathematics
21. Statistical Inference

Unless we are very lucky, the statistics we gather from our small sample will not exactly match the parameters of the population. We can't exactly predict what the population looks like!

Summarising data

We first need to take our data and work out what it is telling us about the *sample*.

We have a number of statistics that we could calculate. First, we look at the *measures of central tendancy*; these tell us, in some sense, what the typical value for a measurement in our sample is.



From *How to lie with statistics*, by Darrell Huff (p.33).

Arithmetic mean

This is the 'usual' average value; it tells us where the 'centre of mass' of the data lies. However, it might not be typical of any actual data value; see the diagram above, or consider that the mean of $\{13, 15, 16, 80\}$ is 31, but no actual data value is anywhere near 31.

Median and quartiles

If we list out all the data values for our measurement in order and then pick the three quartiles (the *lower quartile* is the value that 25% of the values lie below, the *median* is the value that 50% of the values lie below, and the *upper quartile* is the value that 75% of values lie below), then we have some information about the 'shape' of the data. If we have two populations to compare (female versus non-female.)

Mode

The *mode* is the most common value for the measurement.

We also have measures of spread, which tell us how likely we are to find points far away from our central point.

Range and inter-quartile range

These values, the difference between the maximum and minimum and between the two quartiles respectively, give us a rough idea of how spread out our data values are. However, they don't tell us whether the 'typical' data value is close to or far away from the centre — only how far apart the furthest data points are.

Variance and standard deviation

The *variance* s^2 of a set of data is the mean of the squares of the distances of each point to the mean.

$$s^2 = \frac{\sum(x - \bar{x})^2}{n}.$$

The *standard deviation* of the set is s, the square root of the variance.

We will use the central points and the spread measurements of our sample to estimate the same values for the population as a whole.

Drawing an inference

We then need to take our *statistics* and work out how sure we can be about the *parameters* of the population.

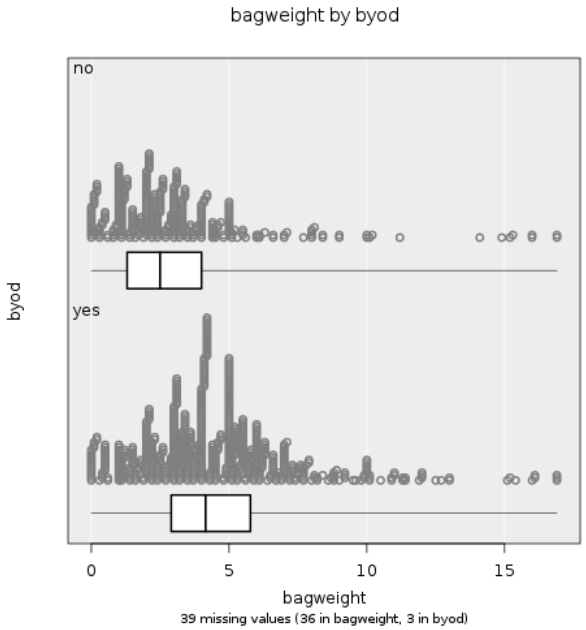
At this level, we won't go into the details too far. What we will do is assume that our median value for our sample is a reasonable guess for the median value of our population, and then write down an *informal confidence interval*:

$$\text{sample median} \pm 1.5 \cdot \frac{\text{IQR}}{\sqrt{n}}$$

(where IQR is the inter-quartile range of the sample). Note: bigger sample size, smaller informal confidence interval. (Why?)

The population median will lie within this range with 90% confidence: if we sample ten times, calculating this value each time, then on average nine of these intervals will contain the population median.*

In order to compare two subsets of a population, it is useful to use side-by-side box and whisker graphs to visualise the shapes of the data sets. For example:



*Next year, we will learn how to make this idea more useful by re-sampling the population a bunch of times to make the confidence intervals we obtain smaller.

Questions

1. Consider a data set with at least three values. Suppose we increase the highest value by 10, and decrease the lowest by 5. Do the mean or the median change? Is it possible for the mode to change?

2. Repeat (1), but now decrease the lowest value by 10 rather than 5 (so the increase of the highest value is the same as the decrease of the lowest).

3. If a data set has an even number of points, is the median ever equal to a value in the set?

4. Consider the numbers $\{2, 3, 5, 5, 5, 7, 9\}$. Which measure(s) of central tendancy (mean, median, or mode) makes sense in the following situations:
   (a) If the numbers represent colours of T-shirts ordered from a website?
   (b) If the numbers represent distances from a point to given destinations?
   (c) If the numbers are survey responses on a scale of 1–10?

5. A set of five data values has a mean of 39. A new data point with value 17 is added to the set; what is the new mean?

6. Compare and contrast the inter-quartile range and the standard deviation as measures of spread. Discuss the following statement: "mean is to median as standard deviation is to range".

7. One important use of the standard deviation is Chebyshev's theorem: for any set of data, and for any constant $k > 1$, the proportion of the data that lies within k standard deviations on either side of the mean is at least $1 - 1/k^2$.
   (a) If a set of data has mean μ and standard deviation σ, at least what proportion of the data lies within the range from $\mu - 3\sigma$ to $\mu + 3\sigma$?
   (b) Show that if a set of data has mean μ and standard deviation σ, then (i) at least 75% of the data lies in the range from $\mu - 2\sigma$ to $\mu + 2\sigma$; (ii) at least 93.8% of the data lies in the range from $\mu - 4\sigma$ to $\mu + 4\sigma$.

8. Consider the following data from the 2017 Census at School. (The question answered was 'How much water did you drink yesterday?'.) Is there any significant difference between the Auckland and Wellington medians?

```
================================================================================
                              iNZight Summary
--------------------------------------------------------------------------------
        Primary variable of interest: region (categorical)
                  Secondary variable: drinkwater (numeric)

        Total number of observations: 1000
   Number omitted due to missingness: 68 (68 in drinkwater)
   Total number of observations used: 932
================================================================================

Summary of drinkwater by region:
--------------------------------

                       Min   25%   Median    75%    Max    Mean      SD   Sample Size
     Auckland Region     0   250      700   1000   4000   799.1   724.7           471
   Wellington Region     0   250      600   1000   4000   774.6   702.1           461


================================================================================
```

9. The following table shows the data collected from a sample of 1000 Census at School respondents (2017 survey), answering the question 'In how many languages can you hold a conversation about a lot of every day things?'. The notation $[a, b)$ means 'all the numbers between a and b, including a but not including b'.

	Female	Male	Totals	Percentages
[1,2)	413	247	660	66%
[2,3)	179	83	262	26.2%
[3,4)	41	18	59	5.9%
[4,5)	10	5	15	1.5%
[5,6)	3	1	4	0.4%
[6,7)	0	0	0	0%
[7,8)	0	0	0	0%
[8,9)	0	0	0	0%
[9,10]	0	0	0	0%
Totals	646	354	1000	100%

   (a) Calculate the statistics for the total sample. Hence write down a guess for the population measurement, and a confidence interval for that guess. Is it more or less than you would expect?

   (b) Calculate the statistics for male and female respondents from the sample separately. Is there a significant difference between your sample medians? Consider the male and female populations of respondents. Can you conclude that, in general, one gender tends to be comfortable with more languages than the other? (Two populations are likely to have different population medians if the informal confidence intervals for the median of each don't overlap.)

10. The following tables show the data collected from a sample of 1000 Census at School respondents answering the question 'What is the main way you usually get to school?'. The first table is sampled from the 2009 data set, and the second from the 2017 data set. Pick a mode of transport, make a hypothesis (with reasoning) as to whether there is likely to be any significant change in usage of your chosen mode over the eight years between the two surveys, and test your hypothesis.

	Totals	Percentages
Bike	51	5.1%
Bus	300	30%
Motor	348	34.8%
Other	8	0.8%
Train	27	2.7%
Walk	266	26.6%
Totals	1000	100%

	Totals	Percentages
Bike	40	4%
Boat	9	0.9%
Bus	298	29.8%
Motor	368	36.8%
Other	18	1.8%
Train	14	1.4%
Walk	253	25.3%
Totals	1000	100%