YuMi Deadly Maths

Year 8 Teacher Resource:

SP – Exploring bias and outliers

Prepared by the YuMi Deadly Centre
Faculty of Education, QUT

QUT

YuMiDeadly
Growing community through education

YuMi Deadly Maths Year 8 Teacher Resource: **SP – Exploring bias and outliers**

ACKNOWLEDGEMENT

We acknowledge the traditional owners and custodians of the lands
in which the mathematics ideas for this resource were developed,
refined and presented in professional development sessions.

TERMS AND CONDITIONS OF YOUR USE OF THE WORK AND RESTRICTED WAIVER OF COPYRIGHT

Prepared by the YuMi Deadly Centre
Queensland University of Technology
Kelvin Grove, Queensland, 4059

research.qut.edu.au/ydc/

Year 8 Statistics and Probability

Exploring bias and outliers

Learning goal	Students will explore the effect of bias and outliers on summary statistics (mean, median, mode, range).
Content description	Statistics and Probability – Data representation and interpretation • Explore the practicalities and implications of obtaining data through sampling using a variety of investigative processes (ACMSP206) • Investigate the effect of individual data values, including outliers, on the mean and median (ACMSP207)
Big idea	Statistics – Construction vs interpretation; inference
Resources	Rope number line, pegs, masking tape, digit cards, elastic, string, calculators, ten strip mat, cards, cubes

Reality

Local knowledge

What does it mean to be biased? Where could there be bias in the local, state or federal environments? Is bias restricted to one age group? Does bias affect everyone? Why? What causes bias?

Prior experience

What are the measures of central tendency that are used in the analysis of data? [mean, median, mode] Check that students are able to calculate the missing statistic:

(a) 8, 3, 6, ?, 4 *What is the missing number if the mean is 5?* [4]

You may need to remind students that the mean is calculated by adding up all the statistics and dividing by the number of statistics. An algebraic statement would help them to work it out:

$$\frac{8+3+6+x+4}{5}=5$$
$$\therefore 21+x=25$$
$$\therefore x=4$$

(b) 4, 7, ?, 11, 16, 19 *What is the missing number if the median is 10?* [9]

(c) 9, 3, 7, 5, 10, ? *What is the missing number if the mode is 7?* [7]

Kinaesthetic

Twelve students were surveyed about how long it took them to get to school, with the following results in minutes:

25, 21, 27, 21, 27, 88, 25, 21, 24, 3, 26, 28

Have students peg cards showing these results in ascending order on the rope number line, or cards may be placed on the floor along a strip of masking tape. (Where there is more than one card with the same number, peg or place it under or beside the identical number.)

Ask questions:

- *What is the range of this sample?* [3 to 88]
- *Where do these numbers cluster?* [in the 20s; put an elastic around these numbers]
- *What do these results tell us?*
- *Where would you expect the mean or average to be?* [in the 20s; place an approximate marker]
- *What may affect the mean?* [extremely low or high numbers, in this case 3 and 88]
- *What name is given to numbers like these?* [outliers]
- *What factors may have impacted these two results?* [distance from school; method of travel]

Cut pieces of string in centimetres to the length of the statistics. Have the students even the lengths up by cutting and joining until each length is the same. Measure the length of the average. Check the mean of the 12 numbers with calculators. [28]

Ask:

- *Is 28 in the centre of most of the numbers in the 20s?* [no]

Have students stand in a line, three deep for 21, one for 24, two for 25, and so on. Ask:

- *What has affected the mean, 28, causing it to be equal to or higher than all the other numbers except one?* [The large 88 has dragged the mean up.]
- *Predict what would happen if 88 was taken out from the sample.* [The mean would drop lower than 28 as the numbers would cluster around the low 20s.]
- *Use your calculators to find the mean excluding 88*. [22.55]
- *Predict what will happen if 88 is left in but 3 is excluded. Will the mean be smaller or larger than 28?* [larger]
- *Use your calculators to find the mean excluding 3 but including 88.* [30.27]
- *Why is the mean now higher than in the other two cases?* [Removing the lowest number in the sample has caused the mean to increase. When both the low outlier (3) and the high outlier (88) were included, the low outlier helped to compensate for the high outlier and kept the mean lower than 30; when the high outlier was excluded but the low outlier was still included, the mean was even lower because the low outlier dragged it downwards.]

Ask students to predict where the mean will be if only the 10 statistics in the 20s are considered:

- *Use your calculators to find the mean of these 10 statistics*. [24.5]
- *Can you give an explanation for this?* [It would be expected that this mean would be between 21 and 28 as all the statistics are now in this range.]
- *Calculate how much has been added to the statistics below 24.5 to make them all equal to 24.5.* [11 – the sum of the differences between the mean of 24.5 and each number below 24.5]
- *What total would you predict needs to be subtracted from the statistics above 24.5?* [11]
- *Check that out with your calculators*.

Using the 12 original statistics on the number line, ask students to find the median:

- *What is the median of the 12 statistics?* [25]
- *Predict the median of the 10 statistics excluding the outliers.* [25] *Why is this so?* [The outliers have been taken from both ends of the sample leaving the median the same.]
- *Would the median be the same if only one of the outliers was removed?* [Yes, in this case, as the two statistics in the centre are the same.]

Using the 12 original statistics on the number line, ask students to find the mode:

- *What is the mode of the 12 statistics?* [21]
- *Is it the same for the 10 statistics excluding the outliers?* [yes]
- *What can you say about the value of the mean, median and mode in this sample?* [The mean was affected and changed considerably because of the inclusion of two outliers, but the median and the mode stayed the same and were not impacted by the outliers.]

Abstraction

Body

A telephone poll of a sample of 100 people living in Australia was conducted to ascertain people's preferred holiday destination. The responses were grouped into two categories: go overseas (O) or stay home in Australia (H).

Telephone Poll

Age	13 years		14 years		15 years		65 years		70 years	
Gender	M	F	M	F	M	F	M	F	M	F
Overseas	3	3	2	6	3	4	2	4	3	5
Home	4	6	5	9	6	10	6	7	6	6
Total	7	9	7	15	9	14	8	11	9	11

M = Male; F = Female

This survey was taken in Brisbane on Wednesday 16 June this year between 3:30 pm and 5:30 pm. The subsequent headline and paragraph in the newspaper read:

Majority of People Remain Home for Holiday

> A recent poll revealed that 65% of people prefer to stay at home in Australia for their holidays. The mean age of 35 years suggests that parents with young children find it too difficult to take small children on overseas trips.

Towards what age group is the poll weighted? What age group has been omitted? Describe the sample size in terms of its validity.

Model a telephone poll conversation between the interviewer (I) and a student (S) for each of the age groups.

I: Good afternoon, I'm taking a poll on where people like to spend their holiday. Would you please tell me whether you stay in Australia or go overseas for your holidays?

S: We stay in Australia.

I: Thank you. Would you tell me your age group – teens, 20s, 30s etc.?

S: Teens.

I: Can I ask how old you are?

S: I'm 13.

I: Thank you for your time.

Student goes and stands on the strip mat in the 13 years square. Repeat process with four other students (one for each age group).

Have boys and girls for each age group stand on the strip mat with cards showing these statistics that represent a telephone poll on whether people travel overseas or stay in Australia for their holidays. Other students sit on the floor around the mat.

Students then place their cards on the strip mat and sit as a class group for the discussion.

- *From the newspaper report, what could have been the cause of the lack of people aged 20 to 60 years responding to the telephone poll?* [They were at work.]
- *What element has caused this distortion or bias to certain age groups?* [Timing: students and retired people were at home at that time to answer the call.]
- *Why is the report inaccurate regarding its statement about families with young children?* [The mean does not reflect the age groups in the sample and no parents with young children were surveyed.]
- *What factors caused bias in this survey and how may these have been avoided?*

Hand Students use Unifix cubes to demonstrate the mean (using compensation), median and mode of given sets of statistics.

Mind Ask students to visualise a set of five statistics that could have a mean, median and mode of 7. Now ask them to see this set of five statistics: 10, 18, 12, 10, 10. *What is the mean, median, mode?*

Creativity Students create their own sample with statistics that are biased by outliers.

Mathematics

Language/ symbols mean, median, mode, sample, range, population, bias, outlier, statistic

Practice

1. Add any two outliers to the following set of statistics so that neither the mean nor the median changes: 42, 42, 42, 42, 43, 43, 43, 43, 44, 44, 44, 44. Explain your reasoning for choosing these numbers.
2. The ages of students on a school bus were as follows:

   10, 12, 16, 12, 11, 13, 14, 12, 11, 14, 12, 12, 15, 13, 12, 16, 14, 10, 12, 11

   (a) Calculate the mean, median and modal age of students on the bus.

   (b) If a 13 year-old student boarded the bus, what could be said about that student's age in comparison to the mean, median and modal age of the other students on the bus?

   (c) If a grandparent boarded the bus, would the mean age change? Explain.

   (d) If a student is selected at random, what is the age most likely to be? Explain.
3. The data appearing in the following tables was obtained from a large high school using 582 students from Year 8 who carried backpacks. The data reflects the colour of backpacks the students carry to school. Groups of students conducted surveys to obtain the data tables that appear below.

Table 1: Data collected from the entire population
(collected by surveying Year 8 students carrying backpacks – 582 students)

	Backpack colour							
Response	**Black**	**Brown**	**Blue**	**Red**	**Yellow**	**Purple**	**Pink**	**Other**
Boys	91	58	66	32	14	9	2	20
Girls	40	15	72	55	20	31	32	25
Totals	131	73	138	87	34	40	34	45

Table 2: Data collected from a random sample
(collected by surveying Year 8 students carrying backpacks, attending period 2 Study Hall – 54 students)

	Backpack colour							
Response	**Black**	**Brown**	**Blue**	**Red**	**Yellow**	**Purple**	**Pink**	**Other**
Boys	12	2	10	6	0	1	0	1
Girls	2	0	8	6	2	2	2	0
Totals	14	2	18	12	2	3	2	1

Table 3: Data collected from a biased sample
(collected by surveying Year 8 girls carrying backpacks and attending period 3 gym class – 42 students)

	Backpack colour							
Response	**Black**	**Brown**	**Blue**	**Red**	**Yellow**	**Purple**	**Pink**	**Other**
Girls	7	0	8	12	4	4	6	1

Table 4: Data collected from a biased sample
(collected by surveying Year 8 boys carrying backpacks and attending period 4 gym class – 36 students)

	Backpack colour							
Response	**Black**	**Brown**	**Blue**	**Red**	**Yellow**	**Purple**	**Pink**	**Other**
Boys	16	2	8	7	1	2	0	0

Students construct bar graphs of the data in the four tables above. Using the tables and graphs, answer the questions below:

1. *Are the results from the entire Year 8 population what you would have expected them to be? Explain.*
2. *What factors might influence a student's choice of colour for a backpack?*
3. *Why do you think that blue was the most popular colour for the entire population?*
4. *How similar is the data collected from the random sample with the entire population? Compare and contrast*.
5. *Would the random sample data in this study be a good predictor for the entire population? Explain*.
6. *Examine the results from the biased samples. Why are these charts labelled as "biased"?*
7. *What are the similarities and differences between the two sets of biased data?*
8. *Would either of the biased sample data in this study be good predictors for the entire population? Explain for both sets of biased data.*
9. *Prepare both a table and a graph that combines the two biased data tables. Would this new table of the combined samples be a good predictor for the entire population? Explain.*
10. *Prepare both a table and a graph that combines the three sample tables (the random and the two biased) to form a larger sample. Would this new table of the larger sample be a good predictor for the entire population? Explain.*

Connections Relate to lists, tables, graphs, percentage, algebra, divergence.

Reflection

Validation Students share and validate their tables and graphs from 9 and 10 above with a partner. They discuss how and where bias can occur in their world context and how this may be avoided.

Application/ problems Provide applications and problems for students to apply to different real-world contexts independently e.g. Develop an unbiased sampling method to survey a school, Years 8–12, where girls form 35% of the school population.

Extension **Flexibility**. Use different models to represent mean, median and mode, employing outliers to demonstrate bias. Ensure students are familiar with many reasons and causes of bias and are able to develop fair data collection strategies.

Reversing. Students are able to move between construction and interpretation of measures of central tendency and can move between recognising and avoiding bias.

Generalising. Outliers differ radically from the majority of statistics, being much higher or lower than others in the sample. The mean is influenced by the inclusion of outliers whereas the median and mode are not as strongly affected by outliers. When conducting surveys and calculating measures of central tendency, factors such as location, timing, sample size and sample type all need to be considered in recognising or avoiding bias and in selecting the most appropriate measure of central tendency.

Changing parameters. Students use algebra to calculate a missing statistic when the mean of the data set is known. Students analyse complex databases that have multiple outliers and complicated bias conditions and use digital technologies.

Teacher's notes

- Ensure that students have a sound understanding of mean, median and mode before introducing outliers. Discuss bias in general situations prior to introducing bias into datasets.
- Students need to be taught the skill of visualising: closing their eyes and seeing pictures in their minds, making mental images; e.g. show a picture of a tennis ball, students look at it, remove the picture, students then close their eyes and see the picture in their mind; then make a mental picture of a soccer ball.
- Suggestions in Local Knowledge are only a guide. It is very important that examples in Reality are taken from the local environment that have significance to the local culture and come from the students' experience of their local environment.
- Useful websites for Aboriginal and Torres Strait Islander perspectives and resources: www.rrr.edu.au; https://www.qcaa.qld.edu.au/3035.html
- Explicit teaching that aligns with students' understanding is part of every section of the RAMR cycle and has particular emphasis in the Mathematics section. The RAMR cycle is not always linear but may necessitate revisiting the previous stage/s at any given point.
- Reflection on the concept may happen at any stage of the RAMR cycle to reinforce the concept being taught. Validation, Application, and the last two parts of Extension should not be undertaken until students have mastered the mathematical concept as students need the foundation in order to be able to validate, apply, generalise and change parameters.

