The Big Earth Data Project

Activities for A level Mathematics

Activity set 1: Atmosphere – The Hole in the Ozone Layer

This set of activities explores the hole in the ozone layer from 1980-2023 featuring daily and monthly data for the ozone hole and various atmospheric and weather measurements. The lessons cover aspects of the statistics element of A level Mathematics.

There is a very accessible guide to the hole in the ozone layer at: discoveringantarctica.org.uk/oceans-atmosphere-landscape/atmosphere-weather-and-climate/the-ozone-hole/

Overview of the lessons

All these lessons start with a section about using statistics in context and about the hole in the ozone layer. The first you use these resources with a class, you may want to use this section fully and then use it more briefly in future lessons.

There are handouts for all the lessons; however, the presentations can be used without the handouts if students are comfortable discussing the charts when displayed on a screen.

All the lessons feature charts produced using Python *notebooks* (interactive coding documents featuring text and code). No knowledge of Python is needed to use the lessons; however, teachers and students can explore the data further using the notebooks linked to at the end of each lesson. See the appendix and the associated presentation for an introduction to using Python for data.

Lesson 1: Diagrams for single variable data

Activity overview:

- Using different diagrams to explore how the size of the ozone hole varies over different seasons and different decades.
- Considering which diagrams are the most and least useful.

Lesson 2: Exploring outliers

Activity overview:

- Use diagrams to explore which values could be considered outliers from a range of ozone hole areas.
- Calculate boundaries for outliers from ozone hole data using the standard conventions.
- Consider two different cases of outliers and discuss whether or not to remove or clean the data of these values.

Lesson 3: Identifying regions in scatter plots

Activity overview:

- Exploring scatter diagrams and finding correlation coefficients in the monthly data to explore the factors associated with ozone hole size.
- Finding regions in scatter plots.
- Exploring the correlation in a region of a scatter plot.

Lesson 1: Diagrams for single variable data

Lesson objectives

- Be able to interpret diagrams for single-variable data, including boxplots, histograms, frequency polygons and cumulative frequency curves.
- Consider the appropriateness of unfamiliar graphs or representations of data.
- Select or critique data presentation techniques in the context of a statistical problem.

Lesson plan:

Understanding the data

- If this is the first time using the context then use the opening section to:
    - Remind students about the importance of understanding the context when working with data.
    - Inform students about the ozone layer and the discovery of the ozone hole. NB the seasons are defined by southern hemisphere month, not solstices/equinoxes.

Exploring the area of the hole in the ozone layer for different seasons

- Students will be shown some charts that demonstrate that the hole in the ozone layer mostly occurs in the southern hemisphere spring. As they look at these they should answer the following questions;
    - What feature of the diagram shows that the hole in the ozone layer mostly occurs in the southern hemisphere spring?
    - Which are the strengths and weaknesses of each diagram?
- Box plot
    - The spring data is further to the right and there's a higher lower quartile, upper quartile and maximum.
    - It's easy to compare medians and quartiles.
    - It only shows five values for the data. You can't tell the size of the data set.
- Strip plot: A dot for each point. They are vertically spaced out to make it easier to read but the vertical position has no meaning.
    - The spring dots are generally further to the right.
    - It's easy to see where the majority of the dots are.
    - For larger data sets (this has 16,071 points split across four categories) it's difficult to interpret. A small proportion of dots is still a lot of dots!
- Separate histograms.
    - The bars for 5 and above are all higher for spring.
    - You can use the histogram to compare the amount of data in each group (it's approximately the same as total area is the same).
    - You can't read averages directly from histograms. They can look very different for different interval widths: you can't tell the other three seasons apart with the minimum interval at 0-5.
- Frequency polygons:
    - How are these similar and how are they different to histograms?
    - What features show that the ozone hole area is mainly observable in the spring?
- How do the diagrams show that most of the time, the hole in the ozone layer occurs in the southern hemisphere spring?

    - For these four diagrams:
        - Box plot
        - Strip plot
        - Split histograms
        - Frequency polygons
    - Which did you find the most useful?
    - Which did you find the least useful?
- For completeness two further diagrams are given: frequency polygons and cumulative frequency. You can discuss these with students too or skip these and move on to the next section.

Exploring the hole in the ozone layer for different decades

- Discuss how the Southern hemisphere spring is where to focus your attention. The diagrams on the slides show the hole in the ozone layer for spring months (September, October, November), grouped by decade.
- In **Task 2** students can use the diagrams to discuss the following questions:
    - What features of the diagrams show a reduction in the size of the hole from 1990 through to 2019?
    - Scientists believe that the hole is still reducing but that there is not enough data yet from 2020s to confirm this. What features of the diagrams show that there is less data from the 2020s?
    - Which diagrams do you find the most useful and which diagrams do you find the least useful?
- Notes:
    - The students should observe that after the hole in the Ozone layer was observed in the 1980s it grew rapidly and by the 1990s was quite large. After steps were taken to reduce CFCs and other chemicals in the atmosphere there has been a steady decline through the 1990s, 2000s and 2010s.
    - The data for the 2020s features 4 years (2020-2023). As this is an incomplete set of data it is difficult to draw conclusions. The first four years of this decade appear to show an increase in the hole; however, without the data for the rest of the decade it is not clear whether this is a random variation or not.
- Ask the students which diagram will be most useful. The following six slides contain full size versions which can be navigated to.

Extension task (using a Python notebook)

Use the notebook at colab.research.google.com/drive/1tTQ8NbZJimjIiozKWNQj0rCp0ZqemGwO

- Explore the other two ozone columns for the different decades. Is it a similar pattern to the ozone hole area?
- Explore the minimum temperature for the different decades. Is there a pattern in this?

Lesson 2: Exploring outliers

Lesson objectives

- Be able to locate outliers in single variable data.
- Consider whether an outlier is an error or a valid piece of data.
- Consider different approaches to dealing with outliers once identified.

Lesson plan

Introduce/remind students of the context

- If this is the first time using the context then use the opening section to emphasise:
    - The importance of using context within statistics.
    - The ozone layer and the discovery of the ozone hole.
- If you have completed lesson with your students, you can skip this section or use it to briefly remind them of the context.

Identifying Outliers

- Define an outlier in general terms. You may want to contrast outliers in Mathematics with anomalies from GCSE Science ('*anomalies*' in GCSE Science are values that don't fit the overall pattern and are almost always discarded). Outliers in mathematics are extreme values which are worthy of investigation but are not always discarded.
- Show the set of data for ozone hole areas and ask students to decide if they would classify any of the points as outliers. Some prompts for discussion:
    - Should the value over 8 million km² be considered an outlier?
    - Should the 4 values between 5 and 6 million km² be considered outliers?
    - Are these likely to be errors or valid, but unusual, values?
    - Remind students that an *outlier* is an **extreme** value that lies outside the pattern of the data.
- Discuss the fact that deciding on outliers by eye isn't always the most sensible way to find points to investigate. There are some conventions that are often used to decide whether a value is an outlier. These are covered in the next section.

Conventions for Outliers

- The following are conventions for determining boundaries for outliers. Different exam boards have slightly different requirements, but all of these are *rules of thumb*. Different multiples of the standard deviation will be used in depending on the context.
    - Using quartiles:
        - More than 1.5 inter-quartile ranges above the upper quartile or below the lower quartile.
    - Using mean and standard deviation:
        - More than 2 standard deviations away from the mean.
        - More than 3 standard deviations away from the mean.
- Use the data for the September Ozone Hole Area to demonstrate calculations for finding the boundaries of outliers. The quartiles are given as percentiles, so 25% is the lower quartiles, 50% is the median and 75% is the upper quartile.
    - The minimum value will be beyond the boundary for the first two definitions above.
    - If you calculate the boundaries for 3 standard deviations, you'll find a boundary that is negative. Given that the measurements are for area, any value below 0 should be considered suspicious.

- Ask your students to calculate boundaries for the ozone hole area for the three months shown on the slide. Note that some of these lower boundaries are negative so mention the point above about negative areas if you haven't already done so.
    - Other than simply recognising outliers by eye for example on a scatter graph, box plots can show outliers by placing dots or crosses for values which are classified as outliers. By convention software will do this for values which are more than 1.5 times the interquartile range above or below the upper and lower quartile.
    - Show your students the box plot for the average ozone hole area in September.

Classifying outliers and cleaning data

- Unlike anomalies in GCSE Science, not all outliers are deleted. Outliers should be investigated to determine whether they are valid data points. If they are not valid then they can be deleted, which is part of cleaning a data set.
- Show the set of box plots to demonstrate the outliers in each of the months.
    - Ask your students whether they would delete the large outlier in the December data. Note that at this stage there isn't enough information to decide.
    - It is possible to examine the data collected around the time of the outlier: The ozone hole area of 8.7 million square kilometres was recorded in 2020.
- Show them the time series to help them decide.
    - In this case it seems reasonable to assume that the December data is a valid data point and would **not** be deleted
- Ask your students to identify an outlier in the stratospheric temperatures for January.
    - Note that with this type of summary data there isn't enough information to know whether there are other values close to the maximum of 0
    - A value of 0°C in the upper atmosphere would be very warm. The value is exactly 0 which seems very unlikely.
    - The maximum value of 0°C is much warmer than any of the other values, you can demonstrate this using the scatter graph.
    - You can also look at the other recorded values for the days around 8th January. It looks as though 'no value' has been replaced with a 0 for some of the data.
    - In addition, the air temperature (taken 2m above the ground) is actually lower than the temperature in the stratosphere (about 25 km above the ground)
    - Ask your students whether they would delete the large outlier in this data.
        - In this case it would seem sensible to delete the entire row.
        - This process is called cleaning the data.
        - Students can investigate this data further using the python notebook.
- Errors can occur in large data sets due to incorrect measurements or recording or even simply when copying a file from location to another. One of the most common ways errors can occur is if a value is missing, it is replaced with a 0.

Summary

- If your exam board requires your students to remember the definitions for the outlier boundaries then this is a good opportunity to ask them to recall them.
- The whole data set and instructions about how to investigate them can be found in the python notebook.

Further exploration (using a Python notebook)

The data and code for creating the charts and tables in this lesson is at:
https://colab.research.google.com/drive/1qQLw7kPow73XC98erzRC0DGtNdjjfV5J?usp=sharing

This notebook can be used to explore outliers in other variables.

Lesson 3: Identifying regions in scatter diagrams

Lesson objectives

- Be able to interpret scatter diagrams for bivariate data.
- Be able to recognise distinct sections of a population within a scatter diagram.
- Use an informal understanding of correlation.

Lesson plan:

Introduce/remind students of the context

- If this is the first time using the context then use the opening section to emphasise:
    - The importance of using context within statistics.
    - The ozone layer and the discovery of the ozone hole.
- If you have completed lesson with your students, you can skip this section or use it to briefly remind them of the context.

Interpreting scatter diagrams and correlation coefficients

- In this activity students will explore whether there is a link between any of the weather or atmospheric measurements and the size of the hole in the ozone layer. Use the times series for the daily data shows that the hole starts to appear in the second half of the year and is usually closed by the end of the year. Discuss how the boxplots confirm this. **The rest of this lesson only uses the data for the southern hemisphere spring (September, October, November).**
- Students complete **Task 1**:
    - Discuss the following questions:
        - What type of correlation is shown in each diagram (positive/negative)?
        - Which show that strongest correlation?
        - Ozone starts to break down when the temperature is below -78°C and there is sufficient sunlight. Are the scatter diagrams consistent with this?
    - Both show negative but the stratosphere temperature is a stronger correlation. This is expected as the ozone layer is in the stratosphere and air temperature is measured at ground level. The ozone layer breaks down when the temperature is below -78°C; when it's above this the hole starts to close, but will still be observable when it's shrinking.

Finding regions is scatter plots

- For task 2 students have If students have difficulty distinguishing between points there are alternative approaches in the appendix of the notebook.
- Students complete **Task 2**:
    - What patterns are there in the diagrams?
    - There are three distinct regions corresponding to different months for the monthly data. This is because the mean daily temperature for each month is different: September is coldest, and November is warmest in the spring in the southern hemisphere. There aren't as defined regions for decade; however, the points for the 1980s are at the bottom as this was when the hole in the ozone layer first developed.

Exploring the correlation in a region of a scatter plot

- The scatter plots and correlation coefficients for ozone hole area and stratosphere temperature for each of September, October and November are given.
- Students complete **Task 3**:
    - Which shows the strongest and which show the weakest correlation?
    - How does this compare to the correlation for the data for the whole of spring?
    - October and November show stronger correlation than the data for the whole of spring. The data for September shows the weakest correlation. Students should be encouraged to look back to the scatter diagram grouped by decade – all the lowest points on the September data are from the 1980s, when the hole was first developing. In the extension task students can explore what the data looks like if the 1980s are removed.

Extension: Identifying regions in the scatter plots for the data since 1990

- Discuss how the scatter plot and box plot relate to each other. Why does this suggest that it might be useful to concentrate on the data from 1990 onwards?
- Use the Python notebook to demonstrate filtering the data for 1990 onwards and finding some scatter diagrams/correlation coefficients.
- Students can use the notebook to explore the strength of the correlation between the hole in the ozone layer and stratosphere temperature for different months.

Further exploration (using a Python notebook)

The data can be explored further using the notebook at:
https://colab.research.google.com/drive/1hPuY9Yv_Dj3JHKwR4BcVadD6SItECHrs

Appendix: Introduction to using Python for data

No prior knowledge of Python is assumed – all the code is given in the notebooks. Further support on using Python for data analysis can be found at: mei.org.uk/introduction-to-data-science/further-support/

Lesson objectives

- Learn how to use Python to explore a big data set.
- Explore the importance of understanding a context to solve a statistical problem.
- Interpret measures of central tendency and variation.

Notebook

Resource: https://colab.research.google.com/drive/1QZyphAr4UhfVhkEEBwDhEve8U8oAs11U

Lesson Plan

Exploring the importance of understanding a context to solve a statistical problem

- Introduce the importance of understanding a context when solving a statistical problem. This is essential to:
    - Clean the data. For example, a value of -1 for temperature in °C is appropriate but -1 for rainfall in mm isn't.
    - Choose appropriate statistics (such as averages) and charts to analyse the data.
    - Interpret the result: i.e. give a meaningful answer to a question.
- Introduction to the ozone context used in this set of lessons on 'Atmosphere'.
    - What is ozone?
    - What is the Ozone layer?
    - Why is it important?
    - Discovering the hole in the ozone layer
    - The development of the hole and the solutions in place.
- Introduction to the data sets
    - How the data was collected with satellites.
    - Information about the columns/variables in the data set. These will become more familiar when students have explored the data.
    - Show the change in the Ozone hole area over the year chart (from the appendix in the activity. Ask the students:
        - What do you notice about when the hole has been observable?

Learning how to use Python to explore a big data set.

- Importance of using technology to work with big data sets. Why use Python:
    - A lot quicker and can handle large data
    - An industry standard
    - Modelling/machine learning tools are built-in
- Introduction using notebooks in either Google Colab:
    - Demo opening the notebook.
    - Save your own copy:
        - Google Colab: *Copy to Drive*

  - Always press ‘Runtime > Run all’ before running any other cells.
  - Demo changing a cell and running it: e.g. change the first code block from 1+1=2.
  - Demo copying and pasting the code into a new block: copy the code to find the grouped statistics and change to ‘month’.
- Students open the notebook and Copy, paste and edit the code to create grouped statistics and charts for months.
- Students use these statistics and charts to answer the **Task 1** questions (either individually or as a group discussion):
  - What season is the hole in the ozone layer mainly visible?
  - Why is the median a more useful average than the mean for summer, autumn and winter?
  - Why are September, October and November listed as ‘spring?’
  - Which months show the biggest values and variation in the hole in the ozone layer? Is mean/std or median/iqr more useful for this?

Interpreting measures of central tendency and variation.

- Ozone depletion: Why does it change over the year?
- Line chart for stratosphere_temperature. This chart shows when the temperature is below -78°C. There is sufficient sunlight from August (1st August is day 213 or 214 of the year). (Code in the appendix of the notebook).
- Students complete **Task 2**
  - Students find the statistics and box plots for stratosphere_temperature by season.
  - Students find the statistics and box plots for stratosphere_temperature by month.
  - Questions:
    - How does the temperature in the stratosphere vary over the year?
    - Did you find the grouping by season or the grouping by month easier to interpret? There is no ‘correct’ answer to this – it is purely a preference!
    - Which average is more appropriate: mean, median or either?
    - Which measure of variation is more appropriate: standard deviation, inter-quartile range or either?
- **Extension task:**
  - Explore how some of the other numerical variables vary over the year.