A Level Maths Support Guide for Edexcel

Large Data Set

CM

Introduction

We will look at the Edexcel Large Data Set (LDS) in this support guide. We will explore the variables in the data set and note key points from it. We will also give some examples of possible exam questions that require you to apply your knowledge of the LDS to illustrate what the expectation is.

What is the Large Data Set?

All exam boards have designed a large data set for the use in Statistics sections of A Level Maths exams. Schools are expected to dedicate some teaching time to exploring the large data set as some exam questions will test knowledge and familiarity of the data set. You will not be required to take copies of the LDS into the exam and you will also not be expected to have a detailed knowledge of the actual data within the data set.

The Edexcel specification states for questions that use the LDS, "the expectation is that these questions should be likely to give a material advantage to students who have studied and are familiar with the data set." In particular, it makes the following remarks about questions testing the LDS:

- questions may assume familiarity with the terminology and contexts of the data and may not explain them. This is so that students that have not seen or studied the data set do not have the same opportunities to access marks as students that have seen and studied the data set;
- questions may use summary statistics or selected data from, or statistical diagrams based on, the data set – these might be given in the question/task, or as stimulus material;
- questions may be based on samples related to the contexts in the data set where students' work with the data set will help them understand the background context;
- questions may require students to interpret data in ways that would be too demanding in an unfamiliar context.

You can download a copy of the data set from the Edexcel website.

What is the Edexcel LDS about?

The Edexcel LDS has samples on weather data in different locations for certain time periods. The data is provided by the Met Office.

The LDS contains the weather data for 5 UK weather stations and 3 weather stations overseas for May–October 1987 and May–October 2015.

The weather stations are:

- UK weather stations: Camborne, Heathrow, Hurn, Leeming and Leuchars
- Overseas weather stations: Beijing, Jacksonville and Perth

The data set contains data for 11 variables:

- Daily Maximum Temperature
- Daily Total Rainfall
- Daily Total Sunshine
- Daily Maximum Relative Humidity
- Daily Mean Windspeed
- Daily Maximum Gust
- Daily Mean Wind Direction
- Daily Maximum Gust Direction
- Cloud cover
- Visibility
- Pressure

We will now explore how the data for each of these variables is presented in the LDS.

Daily Maximum Temperature

Values for the daily maximum temperature are given in Degrees Celsius and to one decimal place. A negative value indicates a reading below 0 °C. If a reading is not available, it is listed as 'n/a'.

Daily Total Rainfall

All the totals given are in millimetres (mm). If the total amount of rainfall collected is less than 0.05 mm, it is referred to as a **trace** of rain. These values are indicated by 'tr' in the data set. If a reading is not available, it is listed as 'n/a'.

Daily Total Sunshine

Values for this are given in hours and to one decimal place. For example, an entry of '4.4 hrs' indicates that there were 4.4 hours of sunshine in that location on that particular day. If a reading is not available, it is listed as 'n/a'.

Daily Maximum Relative Humidity

Relative humidity is a measure of how close the air is to being saturated with water vapour.

Values for this are recorded as percentages (%). Relative humidities above 95% are associated with mist and fog. If a reading is not available, it is listed as 'n/a'.

Daily Mean Windspeed

The daily mean windspeed is given in knots. 1 knot is 1.15 mph. If a reading is not available, it is listed as 'n/a'.

The windspeeds are also categorised according to the Beaufort scale. This is an empirical and discrete scale. You can read a little about the Beaufort scale online.

Daily Maximum Gust

The maximum gust speed is the maximum instantaneous speed that occurred over a 24 hour period.

It is calculated as an average over a 24 hour period. If a reading is not available, it is listed as 'n/a'.

Daily Mean Wind Direction

Two data processes are used to obtain this. The mean direction of the wind is calculated each hour. The value for the daily mean wind direction is then recorded in the LDS as the most frequently recorded (i.e. modal) wind direction of these hourly data captures.

The value is given in degrees relative to the true north. The corresponding cardinal direction is also given.

Daily Maximum Gust Direction

This is the direction the wind was blowing in the hour the corresponding daily maximum gust occurred.

Values are given in degrees relative to the true north. The corresponding cardinal direction is also given.

Cloud cover

This is a discrete variable in the data set. It is measurement of the fraction of the celestial dome covered by cloud.

It is measured in eighths. The technical unit used in this case is called **oktas**.

0 oktas indicates a completely clear sky, while 8 oktas indicates complete overcast.

Visibility

Visibility is measured horizontally. Readings are given in metres. A dash indicates unavailable data.

Pressure

This is recorded in hectopascals (hPa).

The previous unit for measuring pressure was the millibar (mb).

1 bar is 1000 millibars and 1 millibar = 1 hectopascal.

Remarks

It is important to try and remember the different variables involved in the LDS and how they are recorded. There are subtleties associated with the recording of the different variables. Hopefully you can start to see how these subtleties may be draw on in an exam question and give an advantage to students that are aware of these.

Example 1 (From the Edexcel SAMs)

Sara is investigating the variation in the daily maximum gust, t kn, for Camborne in June and July 1987.

She used the large data set to select a sample of size 20 from the June and July data for 1987. Sara selected the first value using a random number from 1 to 4 and then selected every third value after that.

(a) State the sampling technique used by Sara.

(b) From your knowledge of the large data set explain why this process may not generate a sample of size 20.

Comments

This is the first question of the paper. Part (b) is worth 1 mark and requires you to have knowledge that the LDS has gaps because some of the data is not recorded. Clearly, you wouldn't be able to get this mark if you hadn't looked at the LDS in some way.

Example 2 (From the Edexcel SAMs)

Sara was studying the relationship between rainfall, r mm, and humidity $h\%$ in the UK.
She takes a random sample of 11 days from May 1987 for Leuchars from the large data set.

[Some parts of the question omitted]

(e) (i) Comment on the suitability of Sara's sampling method for this study.

(ii) Suggest how Sara could make better use of the large data set for her study.

Comments

The earlier parts of the questions ask you to do some standard processing of the data Sara obtained. While it is in the context of the LDS, you don't need any knowledge of the LDS to answer those parts. However, for these two parts, you do.

Part (e/i) asks us to comment on the suitability of Sara's sampling method for this study. To answer this, we note that Sara was studying the relationship between rainfall and humidity in the UK. Her sample consisted of 11 days in Leuchars from May 1987. This is clearly a very limited sample since it only consists of 11 days from one location and one month – it is unlikely to representative of the whole of the UK across different time periods. Arguably, you can answer this question without seeing the LDS before and just using 'common sense'.

Part (e/ii) does require some knowledge of the LDS. For this part, you need to remember that the LDS contains weather data for more locations and more months. The mark scheme says that a comment such as 'use data from more locations and months' is worth no marks. This is because you have to realise that Sara only wants UK data and then remember that the LDS has data for locations overseas. A comment such as 'use data from more UK locations and months' is suitable for the mark here.

Example 3 (From CM D17 Mock Exam)

Lauren wants to find the average daily mean windspeed in Hurn in 1987.

She only has access to the large data set. She uses it to obtain a simple random sample of the daily mean windspeeds, t knots, on n days in Hurn in 1987.

The data obtained by Lauren is summarised as follows

[Summary statistics given]

[Some parts of the question omitted]

Lauren uses the same sampling method to estimate that the average daily mean windspeed in Hurn in 2015 was 11 mph.

(b) Convert 11 mph into knots.

(c) Hence, compare the average daily mean windspeed in Hurn in 1987 and 2015.

(d) With reference to the large data set, state **one** limitation of your conclusion in part (c).

Comments

This question requires you to have knowledge of the conversion between mph and knots.

For part (d), you need to be aware that the conclusion is limited since the large data set only contains data between May and October and so our comparison may not be valid for the whole year.

Key points to remember

All in all, if you have some basic knowledge about the variables in the LDS, then most exam questions requiring LDS knowledge should be fairly straightforward.

Some key points to remember are:

- there are gaps in the large data set because some values are not available. Unavailable values are given as 'n/a' (or indicated by a dash in the case of visibility)
- the large data set has weather information on 5 UK locations and 3 overseas locations. Be aware of these locations. The large data set also only contains weather information from May–October, so using the data to make conclusions about whole year round weather patterns may not be entirely reliable
- a trace of rainfall indicates a recorded value of rainfall less than 0.05 mm
- conversion between mph and knots
- cloud cover is a discrete variable. It is measured in oktas
- relative humidities above 95% are associated with mist and fog
- it is also important to realise that not all the data variables are available for all the locations. Take note of which variables are available for which locations and in which periods

This is not an exhaustive list of things you need to know and remember. You still need to do some work with the LDS yourself. We still recommend you take the spreadsheet, calculate averages, plot diagrams, look for outliers, make comparisons and so on.

Once you feel confident with the large data set, you can try the 6 practice questions we have created focusing on the LDS. You can find the document on our website under **A Level Learning Resources**.