Topics from GCE AS and A Level Mathematics

|  |
| --- |
| * Interpret histograms for single-variable data, including understanding that area in a histogram represents frequency. Connect to probability distributions.
* Recognise and interpret possible outliers in data sets and statistical diagrams.
* Be able to clean data, including dealing with missing data, errors and outliers.
 |

**Investigation 1**

Based on daily maximum gusts at Heathrow between May and October 2015, what is the expected proportion of days between May and October (inclusive) when there may be delays to passenger flights at Heathrow Airport due to crosswind gusts from and to the north?

**Background**

The speed and direction of wind and wind gusts affect the chance of aircraft taking off safely. So often, when there are strong winds at airports, there are delays to passenger flights.

The crosswind acceptable depends on the specific aircraft being flown.  Typically, single engine aircrafts can't handle winds of more than 10-15 knots.  Medium jets, such as a Boeing737 (an aircraft that can hold up to 215 passengers) will often accept a crosswind as high as 20-25 knots.

At crosswind speeds above 25 knots there will start to be minor delays, depending on airport configuration and aircraft type.  At around 40 knots the delays will start becoming significant.

The runways at Heathrow run west to east (or east to west) so a crosswind would blow from either the north or the south.

For this simplified model consider crosswinds to be maximum gusts blowing from north-north-east (NNE), north (N), north-north-west (NNW), south-south-west (SSW), south (S) and south-south-east (SSE).

**The data**

Open the Excel workbook **Edexceldataset.xlsx.**

1. Select the **Information** worksheet. Read the information in cell **A16.**
2. Which is the more sensible variable to use for this investigation: daily maximum gust or daily mean windspeed?
3. Is the direction of the maximum gust, and hence the cardinal direction, the direction from or to which the wind was blowing?
4. State the target population for this investigation.
5. What type of data are in the database?
6. What type of variable is daily maximum gust?

*Select the worksheet* ***Heathrow May-Oct 2015*** *in the workbook* ***Edexceldataset.xlsx.***

**

It is difficult to analyse these data as they are presented in the dataset. The headers need to be in row 1.



*Copy the data into a new workbook*

*Select the whole worksheet*

*Click on the small blue square in the left hand corner*

*This will select the whole worksheet.*

***Right click*** *then* ***Copy***

*Open a new workbook*

*Select the* ***Office button***

*then* ***New*** *then double click on* ***Blank Workbook***

*Select* ***A1*** *right click* ***Paste***

*Rename the worksheet*

*Double click on the tab* ***Sheet 1*** *at the bottom of the worksheet*

*Type* ***Heathrow 2015,*** *then**click anywhere on the worksheet*

*Delete rows 1 – 5*

*Select rows 1 – 5 right click* ***Delete***

*Save workbook as* ***Heathrow2015***

1. Are there any missing values?

There are several ways to check for missing values.

*Use the COUNTBLANK function in Excel to check for missing values*

*Select* ***P2*** *type* ***=count***

*A list of COUNT functions will appear.*

*Double click on* ***COUNTBLANK***

*This function counts the number of empty cells*

*in a specified range.*

*Select* ***A1*** *to* ***O185*** *and* ***Enter***

*Do not select the whole columns A to O as there*

*are thousands of empty cells in these columns.*

*The result is 0.*

***Save your work***

A problem with this COUNTBLANK method is that missing values could be labelled N/A. \* or have some other coding.

Missing data is coded as n/a (= data not available) in this database.

Another way to find missing values is to sort the data according to the variable you are interested in, either alphabetically or numerically depending on the type of variable, and check the column for coded missing values.

*To sort data in Excel*

*Select the* ***Data*** *tab*

*Select columns* ***A*** *to* ***O***

*Select* ***Sort***

*Tick* ***My data has headers***

*Select* ***Daily Maximum Gust*** *in the* ***Sort by*** *box* ***OK***

*Inspect the data for coded missing values.*

If data are missing do not delete the whole row of data. Either accept the value is missing or, if possible, try to find the missing value.

1. Are there any missing values?
2. Explain why daily maximum gusts from and to the north are being considered in this investigation.
3. Discuss if it is sensible to consider daily maximum gust from the six directions stated.

*Sort the data according to the variable* ***Cardinal Direction for daily maximum gust*** *which is in column* ***O****.*

*Copy headers into* ***sheets 2*** *and* ***3***

*In the worksheet* ***Heathrow 2015*** *select row* ***1*** *right click* ***Copy***

*In worksheet* ***Sheet 2*** *select* ***A1*** *right click* ***Paste***

*In worksheet* ***Sheet 3*** *select* ***A1*** *right click* ***Paste***

*Copy and paste the rows for cardinal directions NNW, N and NNE into* ***Sheet 2***

*In worksheet* ***Heathrow 2015*** *select rows* ***30-31*** *right click* ***Copy****, in worksheet* ***S*heet 2** *select* ***A2*** *right click* ***Paste***

*In worksheet* ***Heathrow 2015*** *select rows* ***37-57*** *right click* ***Copy****, in worksheet* ***Sheet 2*** *select* ***A4*** *right click* ***Paste***

*Rename* ***Sheet 2 North***

*Cells with #### means the cell is not wide enough for the data.*

*Double click on the edge of the column to widen it.*

*Copy and paste the data for cardinal directions SSW, S and SSE into* ***Sheet 3***

*Rename* ***Sheet 3 South.***

***Save your work***

**Investigation 1 (a)**

Based on daily maximum gust at Heathrow between May and October (inclusive) 2015, what is the expected proportion of days between May and October when there may be delays to passenger flights at Heathrow Airport due to crosswind gusts from the south?

Present the findings in an appropriate graph.

**Process**

Plot a histogram to investigate the shape of the distribution for daily maximum gust from the south.

It is tricky to plot histograms in Excel. (See Appendix 1, ‘**Histograms in Excel’,** for instructions on how to do this.) Here the histograms are plotted using GeoGebra.

*Plot a histogram in GeoGebra with class width of 5 knots*

*Open* ***GeoGebra*** *select* ***View*** *then* ***Spreadsheet***

*Copy and paste the data for Daily Maximum Gust for Heathrow 2015 into GeoGebra*

*In the Excel worksheet* ***South*** *select column* ***G (Daily Maximum Gust)***

*Right click* ***Copy***

*In* ***GeoGebra*** *select* ***A1*** *right click* ***Paste***

*Plot a histogram*

*In* ***GeoGebra***

*Select column* ***A***

*Click on the histogram icon*

**

*Click on the  icon select* ***Use Header as Title*** *and* ***Raw Data*** *then* ***Analyse***

*(Note: if the  icon cannot be found then delete the title in A1.)*

**

*Right click on the arrow near the top right hand side of the graph window.*

*Select* ***Set Classes Manually***

*In* ***Start*** *enter* ***0*** *in* ***Width*** *enter* ***5 Enter***

*For this question leave* ***Frequency Type*** *as* ***Count***

*as frequencies will be needed to calculate the proportion of days there may be delays at Heathrow due to crosswinds from the south.*

*It is usual to plot a histogram using frequency density on the vertical axis.*

**

*Scroll down and select* ***Histogram*** *and* ***Frequency Table***

*Check the class intervals are correct*

*Right click select* ***Copy to Graphics View***

*Close the original histogram window.*

*Tidy up the graph and label*

*Right click on* **a = 165** *(this is on the left hand side of the main window)**click on* ***Show Label***

*a=165 will have been removed from the histogram.*

**

*Right click on* **text 1** *(this is on the left hand side of the main window)**click on* ***Show Object***

*The frequency table will have been removed from graph area.*

*Right click on the graph area, not on the bars.*

*Select* ***Graphics*** *(if this does not appear in the menu at first try right clicking in different places in the graph).*

|  |  |
| --- | --- |
| *Select the* ***Basic*** *tab**Set* ***x Min*** *at* ***-1.5****Set* ***y Min*** *at* ***-1*** |  |
| *Select the* ***xAxis*** *tab**Tick* ***Distance*** *enter* ***5****In* ***Label*** *enter****Daily maximum gust (kn)*** |  |
| *Select the* ***yAxis*** *tab**Tick* ***Distance*** *enter* ***2****In* ***Label*** *enter****Frequency*** |  |

*Close the* ***Graphics*** *window*

*Copy the graph into Word*

*Select the* ***Edit*** *tab**then* ***Graphics View to Clipboard***

*In the Word document right click where the histogram is to be pasted* ***Paste***

*Add the title above or below the histogram:*

***Daily maximum gust from the south at Heathrow, May to October 2015***



***Save your work***

1. Comment on the shape of the distribution for the daily maximum gusts from the south at Heathrow, May to October 2015.
2. The two days when the daily maximum gusts were greater than 35 knots would be considered outliers according to the rule for outliers:

 Lower limit = Q1 – 1.5 × IQR

 Upper limit = Q3 + 1.5 × IQR

Give a reason why you would not remove these data from the data set.

1. Based on the data from 2015, what is the proportion of days between May and October when there may be delays to flights at Heathrow due to crosswind gusts from the south?

 $\frac{7}{184}$ = 0.038 = 3.8%

**Report**

**Investigation 1 (b)**

Based on daily maximum gusts at Heathrow between May and October 2015, what is the expected proportion of days between May and October when there may be delays to passenger flights at Heathrow Airport due to crosswind gusts from the north?

Present the findings in an appropriate graph.

**Process**

Using the same scales as for daily maximum gusts from the south, plot a histogram to investigate the shape of the distribution for daily maximum gust from the north.

*(In GeoGebra open a new window before pasting in the data otherwise the second histogram will be plotted on the same grid as the first one.)*

1. Comment on the shape of the distribution for the variable daily maximum gusts from the north at Heathrow, May to October 2015.
2. Based on the data from 2015, what is the proportion of days between May and October when there may be delays to flights at Heathrow due to cross wind gusts from the north?

**Investigation 1**

Based on daily maximum gust at Heathrow between May and October 2015, what is the expected proportion of days between May and October (inclusive) when there may be delays to passenger flights at Heathrow Airport due to crosswind gusts from and to the north?

*Appendix 1- Histograms in Excel*

*To plot a histogram in Excel*

*First make sure the* ***Data Analysis Toolpak*** *is in Excel.*

*Press at the same time* ***Alt T I*** *(i.e. Alt and capital T and capital I)*

*The* ***Add-Ins*** *menu will appear.*

*Tick* ***Analysis ToolPak*** *then* ***OK***

***Data Analysis*** *will appear in the* ***Data*** *tab.*

*In the workbook* ***Heathrow2015*** *select the worksheet* ***South***

*Bins in Excel are the upper boundaries and lower limits of each class.*

*Find the minimum and maximum values of the data to*

*decide on the values of the bins*

*Select* ***P2*** *type* ***Min***

*Select* ***Q2*** *type* ***=MIN(*** *and either select column G or type* ***G:G)*** *then* ***Enter***

*Select* ***P3*** *type* ***Max***

*Select* ***Q3*** *type* ***=MAX(*** *and either select column G or type* ***G:G)*** *then* ***Enter***

*Since the minimum value is 12 and the maximum is 43 use bins 10, 15, 20, … ,45*

*Select* ***R1*** *and type* ***Bins***

*Select* ***R2*** *and enter* ***10***

*Select* ***R3*** *and enter* ***15***

*To copy the pattern 20,25,…,45 in cells R4:R9*

*Select* ***R2:R3*** *take the cursor to the bottom right hand corner of* ***R3*** *until you see a small cross then click and drag Stop at 45 (****R9****)*

*Plot the histogram*

*Select the* ***Data*** *tab**then* ***Data Analysis*** *then* ***Histogram*** *and* ***OK***

*Select* ***Input Range*** *(by clicking into the box) then select the data - column* ***G*** *or* ***G1:G34***

*Select* ***Bin Range*** *then* ***R1:R9***

*Tick* ***Labels*** *so* ***G1*** *and* ***R1*** *will be recognised as headers*

*Select* ***Output Range*** *then select* ***S1***

*Tick* ***Chart Output*** *then* ***OK***

*The output graph is not a histogram as it has gaps and it is not labelled properly.*

*Remove the legend*

*Click on then* ***delete***

*Close the gaps between the columns*

*Right click on a bar*

*Select* ***Format Data Series*** *then* ***Series Options***

*Close* ***Gap width*** *to* ***0 %***

**

*Add a border to the columns*

*Select* ***Border Color*** *then* ***Solid line*** *open* ***Color*** *and choose* ***black*** *and* ***Close***

*Increase the size of the graph*

*Click (you will see a double arrow) and drag the bottom corner of the graph window.*

*Add a title and label the axes*

*Click on* ***horizontal axis (Bins)*** *then type in* ***Daily maximum gust (kn)***  *then* ***Enter***

*The vertical axis is fine.*

*Click on the* ***title (Histogram)*** *then type* ***Daily maximum gust from the south Heathrow May to October 2015*** *then* ***Enter***

*Change the font size of the title to 12.*



*When Excel processes the histogram function it counts how many days are greater than 0 and less than or equal to 10 and enters this count into* ***T2****. This is 0 as the minimum value is 12.*

*The frequency in* ***T3*** *is the number of days when the daily maximum gusts are greater than 10 and less than or equal to 15. The frequency in* ***T4*** *is the number days when the daily maximum gusts are greater than 15 and less than or equal to 20, and so on.*

*This needs to be shown on the x-axis scale. A number in brackets shows this is a limit and that the exact value is not included in the column.*

Correct the scale on the horizontal axis

*Select* ***S2*** *and type* ***<10*** *then* ***Enter***

*Select* ***S3*** *and type (****10)-15*** *then* ***Enter*** *and so on*

*Select* ***S10:T10*** *and* ***delete***

*Add horizontal major and minor gridlines to help read the frequencies*

*Right click on the vertical axis*

*Click on* ***Add Major Gridlines*** *and* ***Add Minor Gridlines***



***Save your work***

Note:

* The labelling of the *x*-axis is a problem in Excel. This should be displayed as a scale.
* To show frequency density on the vertical scale divide the frequency by the class width in
* The shape of the histogram may vary due to the boundaries and limits used by different software packages.