Higher Applications of Maths Revision Booklet

Statistics Unit



Topic List

1. Data and Sampling

2. Graphical and Numerical Representations

3. Normal Distributions and Histograms

4. Correlation

5. Linear Regression

6. Hypothesis Testing

Some questions will require access to R studio. If you do this at home remember you can do so through Glow and the tile ‘Noteable’.

**Section 1: Data and Sampling**

1. Give ***two*** words to describe the data below.

(a) Height of a tree.

(b) Genus of tree.

(c) How much money a school spends in a year.

(d) The number of staff employed.

(e) Population of a country.

(f) How many students pass a test.

(g) What subjects a student does.

(f) What position a player plays in.

(h) The price of bread in a shop.

(i) The weight of a salmon.

(j) Cleanliness ratings from 1 to 5.

2. A survey is being done about S1’s experience in the first few weeks of term. A teacher asks their class. What in this example is the population and what is the sample?

3. A survey is done in a shop where they ask every 10th customer to fill in. What kind of survey is this?

4. If you asked all of S3 what the food in the canteen was like, what kind of sample is this?

5. In a supermarket there are 80 Men employed and 120 Women. If you wanted to make a stratified sample of 40 employees how many of each gender should you ask?

6. In a university there are 500 First years, 400 Second Years, 350 Third Year and 250 Fourth Years. If you did a stratified sample of 90 students, how many from each sub category should you ask?

**Section 2: Numerical and Graphical Representation**

From the following list of graphs and charts

1. Bar Chart
2. Line Graph
3. Pie Chart
4. Box Plot
5. Scatter Plot
6. Stem and Leaf Diagram

1. State a type of graph that would be suitable to display.

1. The amount of people who voted for different political parties
2. People’s favourite genre of movie
3. The temperature throughout the week
4. Pupil’s weight against their height
5. How people get to work
6. A shop’s sale figures throughout the month of June
7. The temperature against how many people go to the cinema

2. Calculate the mean and standard deviation of the following

60, 45, 72, 90, 83

3. The following graph shows show people get to work.



If 500 people were asked then how many said bus?

**Section 3: Normal Distribution and Histograms**

1. What 3 things can you to test if data is normally distributed?

2. Draw the following Histogram and state if it is normally distributed or skewed.

|  |  |
| --- | --- |
| Interval | Frequency |
| 130 ≤ x < 140 | 2 |
| 140 ≤ x < 145 | 5 |
| 145 ≤ x < 150 | 15 |
| 150 ≤ x < 160 | 8 |
| 160 ≤ x < 175 | 9 |

3. Draw the following Histogram and state if it is normally distributed or skewed.

|  |  |
| --- | --- |
| Interval | Frequency |
| 0 ≤ £ < 20 | 40 |
| 20 ≤ £ < 30 | 50 |
| 30 ≤ £ < 40 | 55 |
| 40 ≤ £ < 50 | 40 |
| 50 ≤ £ < 100 | 50 |

4. Using the method discussed in question 1. Show if any of the following are normally distributed.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 24 | 29 | 30 | 32 | 45 | 31 | 22 | 28 | 29 | 52 | 34 | 42 |
| Height | 177 | 172 | 176 | 178 | 183 | 180 | 179 | 178 | 185 | 175 | 176 | 180 |
| IQ | 99 | 102 | 110 | 105 | 104 | 98 | 97 | 105 | 125 | 114 | 106 | 104 |
| Time to complete 1km | 350 | 280 | 260 | 325 | 310 | 290 | 300 | 330 | 360 | 390 | 360 | 320 |

Is there enough information to tell us if these variables are actually normally distributed.

**Section 4: Correlation**

1. State what each of the following correlation coefficients means.

(a) r = -0.8

(b) r = 0

(c) r = -0.28

(d) r = 1

(e) r = 0.005

(f) r = 0.62

2. Input the following tables into R studio. Calculate the correlation coefficient and state what that means.

(a)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 2002 | 2004 | 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2018 |
| Number of Cats | 70 | 46 | 57 | 78 | 32 | 67 | 90 | 45 | 40 |

(b)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Number of Bees Spotted (millions) | 4.5 | 6 | 2.5 | 1.1 | 1.9 | 4.7 | 5.2 | 3.1 | 5 | 7.8 | 1.1 | 0.75 |
| Amount of Honey Produced (tonnes) | 600 | 700 | 400 | 190 | 200 | 650 | 710 | 450 | 715 | 900 | 140 | 90 |

(c)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Number of cars on the road (millions) | 18 | 18.4 | 18.9 | 19.3 | 19.5 | 19.8 | 20.5 | 20.9 | 21.1 | 21.5 | 21.9 | 22.6 |
| Average price of petrol that year  | 1.34 | 1.36 | 1.4 | 1.45 | 1.40 | 1.51 | 1.32 | 1.58 | 1.77 | 1.68 | 1.70 | 1.8 |

**Section 5: Linear Regression**

1. Look at the data below and input it into R Studio

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Trees (thousands) | 925 | 882 | 750 | 640 | 590 | 540 | 420 | 390 | 350 |
| CO2 Levels | 3.2 | 4.5 | 5.7 | 6.2 | 8.1 | 9.7 | 11.5 | 14.1 | 19.9 |

(a) Calculate the correlation coefficient between the number of trees and the CO2 levels.

(b) Create an equation of linear regression and state it in the form y = mx + c

(c) Create a line of best fit on R Studio

(d) Make a prediction about the CO2 levels if there were 200 (Thousand) trees.

2. Look at the data below and input into R Studio.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Viewers of CL Final(millions) | 305 | 309 | 312 | 308 | 315 | 330 | 350 | 317 | 360 |
| No. of Amateur football teams(thousands) | 1.25 | 1.28 | 1.31 | 1.32 | 1.35 | 1.38 | 1.39 | 1.44 | 1.45 |

(a) Calculate the correlation coefficient between the viewers of the champions league final and the number of amateur teams.

(b) Calculate the equation of linear regression and state in the form y = mx + c

(c) Create a line of best fit on R studio

(d) Make a prediction about the number of Amateur football teams if the viewers of the champions league final was 400 (million).

3. Look at the data below and input into R studio.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Distance from a volcano (miles) | 1.9 | 2.5 | 1.6 | 5.8 | 4.2 | 3.9 | 3.3 | 2.8 | 1.7 | 1.9 | 0.75 | 1.4 | 5.6 | 4.7 |
| Amount of Tsunamis  | 17 | 15 | 16 | 13 | 12 | 11 | 12 | 10 | 12 | 14 | 10 | 13 | 15 | 14 |

(a) Calculate the correlation coefficient between the distance from a volcano and the amount of Tsunamis.

(b) Calculate the equation of linear regression and state in the form y = mx+ c

(c) Create a line of best fit on R Studio

(d) Make a prediction about the number of Tsunamis that will have in a town 5.5 miles away from a volcano.

**Section 6: Hypothesis Tests**

1. A gym record information about it’s customers and how often they go in the table below?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Age 16-24 | 15-30 | 30+ |
| Once a week. | 70 | 110 | 308 |
| Between 2 and 4 times a week. | 152 | 280 | 452 |
| More than 4 times a week. | 8 | 29 | 3 |

(a) What kind of data is this information and what kind of test does it suggest we will use?

(b) State the null and alternative hypothesis.

(c) Perform a statistical test and interpret these results.

2. Information about a drug being tested is below. There are two columns. The control group are not given the drug but instead a placebo, while the treatment are given a drug. Data about their average resting heart rate is included below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Control | 90 | 81 | 124 | 110 | 96 | 91 | 82 | 117 | 109 | 84 | 72 | 84 |  |
| Treatment | 110 | 112 | 98 | 128 | 124 | 118 | 119 | 120 | 82 | 95 | 100 | 103 | 75 |

(a) State the null and alternative hypothesis.

(b) Assuming the data is not normally distributed state fully which kind of test you are going to use.

(c) Perform this test and interpret your results.

(d) How could this information be used?

3. The amount of pollution in 8 cities across the UK is tracked. A program is produced to lower pollution slowly across 10 years, results are recorded below.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F | G | H |
| Year 0 | 12.3 | 11.1 | 15.6 | 13.5 | 14.2 | 16.3 | 17.3 | 14.2 |
| Year 10 | 10.7 | 11.2 | 9.4 | 8.5 | 10.5 | 13.2 | 14.1 | 10.8 |

(a) State the null and alternative hypothesis

(b) Show if the data is suitable for a t test.

(c) Perform a hypothesis test and interpret your results.

4. The height that people can jump in 2 gyms is recorded.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gym 1 | 58 | 62 | 71 | 64 | 69 | 65 | 68 | 62 | 79 | 72 | 85 | 83 | 55 | 64 | 75 | 54 | 59 | 62 | 64 |
| Gym 2 | 60 | 54 | 55 | 57 | 67 | 82 | 75 | 74 | 72 | 76 | 81 | 72 | 66 | 65 | 68 | 70 | 82 | 75 | 68 |

Stating the null and alternative hypothesis, perform statistical analysis to show if there is anything to show if people who can jump higher go to one gym or the other.