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. Numerical and Continuous

(b) Genus of tree. Categorical and Nominal

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

(d) The number of staff employed. Numerical and Discrete

(e) Population of a country. Numerical and Discrete

(f) How many students pass a test. Numerical and Discrete

(g) What subjects a student does. Categorical and Nominal

(f) What position a player plays in. Categorical and Nominal

(h) The price of bread in a shop. Numerical and Discrete

(i) The weight of a salmon. Numerical and Continuous

(j) Cleanliness ratings from 1 to 5. Categorical and Ordinal

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? Population is all of S1 and the class 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? Systematic Sampling

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

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?

80 + 120 = 200

80/200 x 40 = 16

120/200 x 40 = 24

16 Men and 24 Women

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?

Total = 1500

First Years 500/1500 x 90 = 30

Second Years 400 / 1500 x 90 = 24

Third Years 350 / 1500 x 90 = 21

Fouth Years 25 / 1500 x 90 = 15

**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 Bar Chart/Pie Chart
2. People’s favourite genre of movie Bar Chart/Pie Chart
3. The temperature throughout the week Line Graph
4. Pupil’s weight against their height Scatter Plot
5. How people get to work Bar Chart/Pie Chart
6. A shop’s sale figures throughout the month of June Line Graph
7. The temperature against how many people go to the cinema Scatter Plot

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?

40% of 500 = 200

**Section 3: Normal Distribution and Histograms**

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

Calculate the median and mode and see if they are close together

Create a histogram and see if the data is in a bell shape

Perform a Shapiro Test and see if p > 0.05

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 |

**Age:**

> Age<-c(24,29,30,32,45,31,22,28,29,52,34,42)

> summary(Age)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 22.00 28.75 30.50 33.17 36.00 52.00

#Mean and median are relatively close together.

> hist(Age)

#Histogram does not appear to be bell shaped and instead looks like it is skewed to the right.

> shapiro.test(Age)

 Shapiro-Wilk normality test

data: Age

W = 0.89852, p-value = 0.1518

#Since p is >0.05 this implies the data is in fact normally distributed.

**Height:**

> summary(Height)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 172.0 176.0 178.0 178.3 180.0 185.0

#Mean and median are relatively close together.

> hist(Height)

#Histogram does not appear to be bell shaped as the most results are not in the middle.

> shapiro.test(Height)

 Shapiro-Wilk normality test

data: Height

W = 0.97943, p-value = 0.9629

#Since p is >0.05 this implies the data is in fact normally distributed.

**IQ:**

> summary(IQ)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 97.0 101.2 104.5 106.6 111.0 125.0

#Mean and median are relatively close together.

> hist(IQ)

#Histogram does not appear to be bell shaped and instead looks like it is skewed to the right.

> shapiro.test(IQ)

 Shapiro-Wilk normality test

data: IQ

W = 0.90871, p-value = 0.2054

#Since p is >0.05 this implies the data is in fact normally distributed.

**OneKm:**

> summary(OneKm)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 260.0 297.5 322.5 325.4 352.5 390.0

#Mean and median are relatively close together.

> hist(OneKm)

#Histogram does not appear to be bell shaped, instead

> shapiro.test(OneKm)

 Shapiro-Wilk normality test

data: OneKm

W = 0.9632, p-value = 0.8283

# Since p >0.05 this does imply the data is normally distributed.

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

Most likely not. Shapiro tests are unreliable unless there is a substantial amount of data.

**Section 4: Correlation**

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

(a) r = -0.8 Strong Negative Correlation

(b) r = 0 No correlation

(c) r = -0.28 Weak negative correlation

(d) r = 1 Perfect positive correlation

(e) r = 0.005 No correlation

(f) r = 0.62 moderate positive correlation

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 |

> Year<-c(2002,2004,2006,2008,2010,2012,2014,2016,2018)

> Cats<-c(70,46,57,78,32,67,90,45,40)

> cor.test(Year,Cats)

 Pearson's product-moment correlation

data: Year and Cats

t = -0.43243, df = 7, p-value = 0.6784

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 -0.7455568 0.5631460

sample estimates:

 cor

-0.1613024

Weak Negative Correlation

(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 |

> Bees<-c(4.5,6,2.5,1.1,1.9,4.7,5.2,3.1,5,7.8,1.1,0.75)

> Honey<-c(600,700,400,190,200,650,710,450,715,900,140,90)

> cor.test(Bees,Honey)

 Pearson's product-moment correlation

data: Bees and Honey

t = 15.588, df = 10, p-value = 2.414e-08

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 0.9281946 0.9945560

sample estimates:

 cor

0.9800376

Strong/Almost Perfect positive correlation

(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 |

> CarsRoad<-c(18,18.4,18.9,19.3,19.5,19.8,20.5,20.9,21.1,21.5,21.9,22.5)

> Petrol<-c(1.34,1.36,1.4,1.45,1.4,1.51,1.32,1.58,1.77,1.68,1.70,1.8)

> cor.test(CarsRoad,Petrol)

 Pearson's product-moment correlation

data: CarsRoad and Petrol

t = 5.3292, df = 10, p-value = 0.0003334

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 0.5648889 0.9600562

sample estimates:

 cor

0.8599899

Strong positive correlation

**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.

> CO2Levels<-c(3.2,3.5,5.7,6.2,8.1,9.7,11.5,14.1,19.9)

> cor(Tree,CO2Levels)

[1] -0.9173174

Strong Negative Correlation

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

> lm(CO2Levels~Tree)

Call:

lm(formula = CO2Levels ~ Tree)

Coefficients:

(Intercept) Tree

 23.56549 -0.02373

Co2 Levels = -0.02373(Trees) + 23.56549

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

plot(Tree,CO2Levels)

abline(lm(CO2Levels~Tree))



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

> predict(lm(CO2Levels ~ Tree), newdata=data.frame(Tree=200))

 1

18.82011

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.

> CLFinal<-c(305,309,312,308,315,330,350,317,360)

> AFTeams<-c(1.25,1.28,1.31,1.32,1.35,1.38,1.39,1.44,1.45)

> cor(CLFinal,AFTeams)

[1] 0.7472319

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

> lm(AFTeams~CLFinal)

Call:

lm(formula = AFTeams ~ CLFinal)

Coefficients:

(Intercept) CLFinal

 0.509733 0.002609

No. of Amateur teams = 0.002609(Viewers of CL final) + 0.509733

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

> plot(CLFinal,AFTeams)

> abline(lm(AFTeams~CLFinal))



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

> predict(lm(AFTeams~CLFinal), newdata=data.frame(CLFinal = 400))

 1

1.553422

1553 teams

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.

> Volcanomiles<-c(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)

> Tsunamies<-c(17,15,16,13,12,11,12,10,12,14,10,13,15,14)

> cor.test(Volcanomiles,Tsunamies)

 Pearson's product-moment correlation

data: Volcanomiles and Tsunamies

t = 0.11901, df = 12, p-value = 0.9072

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 -0.5054518 0.5548080

sample estimates:

 cor

0.03433595

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

> lm(Tsunamies~Volcanomiles)

Call:

lm(formula = Tsunamies ~ Volcanomiles)

Coefficients:

 (Intercept) Volcanomiles

 13.00580 0.04563

No. of Tsunamis = 0.04563(Miles away from a Volcano) + 13.00580

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

plot(Volcanomiles,Tsunamies)

abline(lm(Tsunamies~Volcanomiles))



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

> predict(lm(Tsunamies~Volcanomiles), newdata=data.frame(Volcanomiles=5.5))

 1

13.25677

**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?

They are categorical which implies we use a Chi Square Test.

(b) State the null and alternative hypothesis.

Ho - There is no relationship between how old someone is and how often they go to the gym.

H1 – There is a relationships between how old someone is and how often they go to the gym.

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

GymAges<-matrix(c(70,110,308,152,280,452,8,29,3), nrow = 3, byrow=TRUE)

rownames(GymAges)<-c("Once a Week", "Between 2 and 4 times a week", "More than 4 times a week")

colnames(GymAges)<-c("Age 16-24", "15 - 30", "30+")

GymAges

chisq.test(GymAges)

Pearson's Chi-squared test

data: GymAges

X-squared = 61.053, df = 4, p-value =

1.743e-12

Since p < 0.05 we ject the null hypothesis and conclude that there is a relationship between someone’s age and how often they go to the gym.

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.

Ho – there is no difference in the heart rates of people on the control drug or the treatment drug

H1 – There is a difference in the heart rate of people on the control drug and the treatment drug

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

A 2 sample T Test (unpaired)

(c) Perform this test and interpret your results.

> t.test(Control,Treatment)

 Welch Two Sample t-test

data: Control and Treatment

t = -1.7636, df = 22.805, p-value = 0.09119

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -24.912005 1.988928

sample estimates:

mean of x mean of y

 95.0000 106.4615

Since p > 0.05 this means we cannot reject the null hypothesis and conclude that there might be no difference in the heart rates of people on the control drug or the treatment drug.

(d) How could this information be used?

To test if the drug is safe for people to use if they have a heart condition.

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

Ho – There is no difference in the amount of pollution 10 years apart.

H1 – There is a difference in the amount of pollution 10 years apart

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

> summary(Year0)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 11.10 13.20 14.20 14.31 15.78 17.30

> summary(Year10)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 8.50 10.22 10.75 11.05 11.70 14.10

> hist(Year0)

> hist(Year10)

> shapiro.test(Year0)

 Shapiro-Wilk normality test

data: Year0

W = 0.97979, p-value = 0.9619

> hist(Year10)

> shapiro.test(Year10)

 Shapiro-Wilk normality test

data: Year10

W = 0.94101, p-value = 0.6211

Since mean and median is close together, the histograms are roughly bell shaped and both the shapiro wilks tess p values are > 0.05 then we can use a paired t test.

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

> t.test(Year0,Year10, paired=TRUE)

 Paired t-test

data: Year0 and Year10

t = 4.7918, df = 7, p-value = 0.001985

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 1.652559 4.872441

sample estimates:

mean of the differences

 3.2625

Since p < 0.05 then we can reject then ull hypothesis and conclude that there is a difference in the amount of pollution before and after the 10 years.

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.

> summary(Gym1)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 54.00 62.00 64.00 66.89 71.50 85.00

> summary(Gym2)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 54.00 65.50 70.00 69.42 75.00 82.00

> hist(Gym1)

> hist(Gym2)

> shapiro.test(Gym1)

 Shapiro-Wilk normality test

data: Gym1

W = 0.94469, p-value = 0.3197

> shapiro.test(Gym2)

 Shapiro-Wilk normality test

data: Gym2

W = 0.95048, p-value = 0.4026

Since the data is normally distributed we are going to use a 2 sample t test

> t.test(Gym1,Gym2)

 Welch Two Sample t-test

data: Gym1 and Gym2

t = -0.89607, df = 35.971, p-value = 0.3762

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -8.244334 3.191702

sample estimates:

mean of x mean of y

 66.89474 69.42105

Since p > 0.05 this means we reject the null hypothesis and conclude that in all likelihood there is no difference in the amount people can jump in either Gym.