A Level · Edexcel · Maths

4 hours 30 questions

# 2.5 Further Correlation & Regression (A Level only)

2.5.1 PMCC & Non-linear Regression / 2.5.2 Hypothesis Testing for Correlation

| | |
|---|---|
| Easy (9 questions) | /69 |
| Medium (7 questions) | /40 |
| Hard (7 questions) | /50 |
| Very Hard (7 questions) | /66 |
| **Total Marks** | **/225** |

**Scan here to return to the course**
or visit savemyexams.com


© 2024 Save My Exams, Ltd.

# Easy Questions

**1 (a)** Explain what is measured by the Pearson product moment correlation coefficient.

**(2 marks)**

**(b)** The product moment correlation coefficient between two variables is denoted $r$. Five different values of $r$, rounded to four decimal places, are given below:

$$r_1 = 0.0000$$
$$r_2 = 0.9812$$
$$r_3 = -1.0000$$
$$r_4 = 0.7652$$
$$r_5 = -0.7098$$

Match each of the following four scatter graphs, showing observations from different bivariate data sets, to one of the values of $r$ given above. You should use each given


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

value of $r$ no more than once.

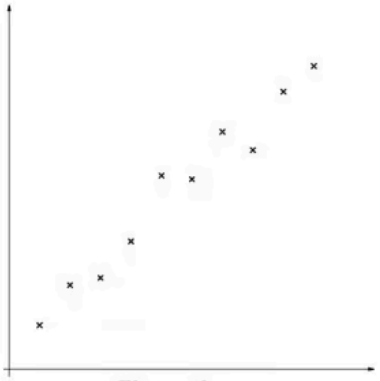


Figure 1

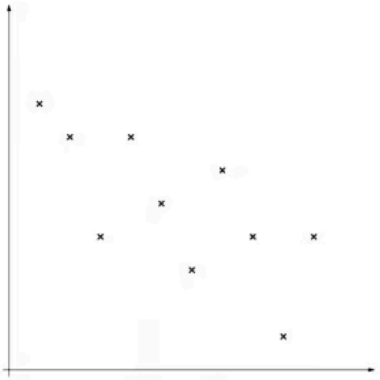


Figure 2

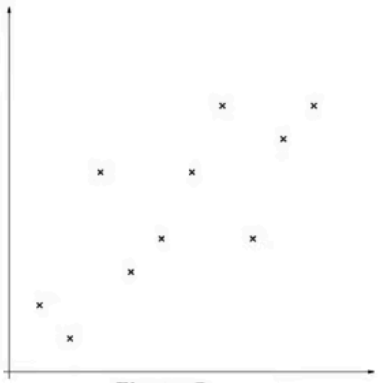


Figure 3

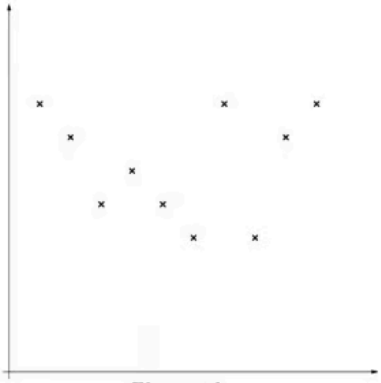


Figure 4

**(4 marks)**

**(c)** Sketch a scatter graph for the remaining value of $r$ from the list above.

**(2 marks)**

**2** Write suitable null and alternative hypotheses for each of the following situations.


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

(i) A recording studio is interested in whether the increasing age of a band's lead singer decreases the number of records the band will sell.

(ii) A researcher for an online gaming company believes that the higher the number of free revivals available in a game, the more time people will spend playing the game.

(iii) A beach umbrella manufacturer is carrying out a test to see if there is correlation between temperature and the number of beach umbrellas sold.

(iv) The developer of a new cryptocurrency tests, at the 5% level of significance, for any correlation between the new cryptocurrency's net value and that of a more popular cryptocurrency. She calculates the product moment correlation coefficient between the two cryptocurrency's net values to be $r = -0.3452$.

**(4 marks)**


© 2024 Save My Exams, Ltd.

**3 (a)** For the following null and alternative hypotheses, state whether the test is a one-tailed or a two-tailed test and write down the type of correlation that is being tested for.

(i) $H_0 : \rho = 0,\ H_1 : \rho > 0.$

(ii) $H_0 : \rho = 0,\ H_1 : \rho \neq 0.$

(iii) $H_0 : \rho = 0,\ H_1 : \rho < 0.$

**(3 marks)**

**(b)** The table below gives the critical values, for different significance levels, of the product moment correlation coefficient, *r*, for a sample of size 30.

| One tail | 10% | 5% | 2.5% | 1% | 0.5% | One tail |
|---|---|---|---|---|---|---|
| Two tail | 20% | 10% | 5% | 2% | 1% | Two tail |
| | 0.2407 | 0.3061 | 0.3610 | 0.4226 | 0.4692 | |

For each set of hypotheses in part (a), use the table above to determine the critical region for a hypothesis test at the 10% level of significance for a sample of size 30.

**(3 marks)**


© 2024 Save My Exams, Ltd.

**4 (a)** It is claimed that there is **negative** correlation between two variables $x$ and $y$. A hypothesis test is carried out to test the claim and the null hypothesis is given as $H_0 : \rho = 0$.

(i) Describe what a null hypothesis of $\rho = 0$ means about the relationship between $x$ and $y$.

(ii) Describe what a negative correlation would suggest about the relationship between $x$ and $y$.

(iii) State a suitable alternative hypothesis to test for negative correlation between $x$ and $y$.

**(3 marks)**

**(b)** The critical value for this hypothesis test is found to be $-0.3674$.

(i) Explain what is meant by a critical value, within the context of hypothesis testing.

(ii) Write down the critical region for this hypothesis test.

**(2 marks)**

**(c)** The product moment correlation coefficient, $r$, between these two variables is calculated to be $r = -0.3175$.

Explain the difference between the statistic, $r$, and the parameter, $\rho$.

**(1 mark)**

**(d)** By comparing the test statistic with the critical value, conclude the hypothesis test.


© 2024 Save My Exams, Ltd.

**(1 mark)**


© 2024 Save My Exams, Ltd.

**5 (a)** Pim collects data on the amount of time she can hold plank each morning, $t$ minutes, and the amount of sleep, $s$ hours, she got the night before.

| Amount of sleep, $s$ hours | 6.21 | 8.15 | 7.52 | 7.19 | 6.18 | 5.28 | 9.03 | 6.01 | 7.55 | 8.39 |
|---|---|---|---|---|---|---|---|---|---|---|
| Time holding plank, $t$ mins | 0.92 | 1.13 | 1.07 | $x$ | 0.99 | 0.96 | 1.12 | 0.98 | 1.20 | 1.09 |

The product moment correlation coefficient for these data is calculated as $r = 0.7536$.

Pim plots this information on a scatter graph and draws, by eye, a line of best fit.

(i) Describe the correlation between $s$ and $t$.

(ii) State, with a reason, whether Pim's line of best fit should have a positive or a negative gradient.

**(2 marks)**

**(b)** Pim calculates the equation of the regression line of $t$ on $s$ to be $t = 0.08s + 0.45$.

(i) Using the regression line, estimate the value of $x$ in the table above and explain why it is only an estimate.

(ii) Give an interpretation of the value 0.45 in the equation of the regression line.

(iii) Give an interpretation of the value 0.08 in the equation of the regression line.

**(4 marks)**


© 2024 Save My Exams, Ltd.

**(c)** Pim says that if she sleeps for 13 hours then she will be able to hold plank for roughly 1.5 minutes. Give two reasons why Pim's claim could be incorrect.

**(2 marks)**

**(d)** One morning Pim can hold plank for one minute. Explain why the regression line should not be used to predict how long Pim slept the night before.

**(1 mark)**


© 2024 Save My Exams, Ltd.

**6 (a)** Andy, a preschool teacher, is exploring whether a new 'Mindfulness for Toddlers' course is helping the children to learn quicker. Andy devises a test where he times how long the nine toddlers in his class can sit in meditation and how long it takes them to solve a simple puzzle afterwards.

The table below shows the amount of time, $m$ to the nearest minute, each child spent meditating and how long, $p$ minutes, it took them to solve the puzzle afterwards.

| $m$ | 5 | 4 | 2 | 10 | 3 | 5 | 1 | 2 | 4 |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | 2.8 | 3.6 | 4.5 | 1.8 | 5.1 | 2.8 | 7.0 | 8.0 | 2.5 |

Andy suspects that there is an exponential relationship between the times so he decides to code the data using the changes of variable $X = m$ and $Y = \ln p$.

Complete the table below for $X$ and $Y$, giving each value of $Y$ to two decimal places.

| $X$ | 5 | 4 | 2 | 10 | 3 | 5 | 1 | 2 | 4 |
|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 1.03 | 1.28 | 1.50 | 0.59 | 1.63 | 1.03 | | | |

**(3 marks)**

**(b)** Andy calculates the product moment correlation coefficient for the relationship between $m$ and $p$ to be $r_1 = -0.772$ and between $X$ and $Y$ to be $r_2 = -0.862$.

State, giving a reason, whether there is stronger correlation between $m$ and $p$ or between $X$ and $Y$.

**(2 marks)**

**(c)** Andy calculates the equation of the regression line of $Y$ on $X$ to be $Y = 1.98 - 0.162X$. A new student joins the class and spends 4 minutes meditating.

(i) Write down the corresponding value of $X$.

(ii) Use the regression line to estimate the value of $Y$.

(iii) Hence estimate how long it takes the student to solve the puzzle.

**(3 marks)**

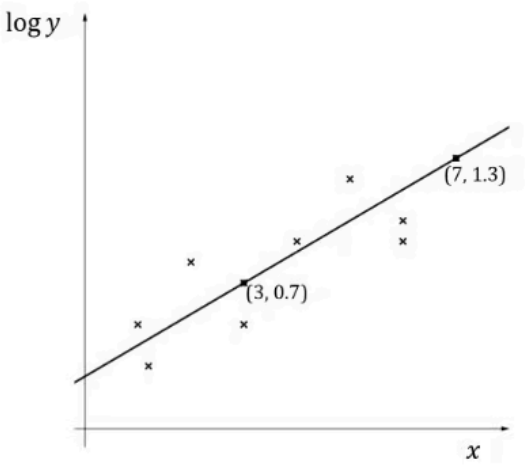

© 2024 Save My Exams, Ltd.

**7 (a)** It is believed that the relationship between two variables, $x$ and $y$, can be modelled by $y = bp^x$.

By first taking logarithms of both sides and then by using the laws of logarithms, show that $y = bp^x$ can be written as $\log y = \log b + x \log p$.

**(2 marks)**

**(b)** The scatter diagram below shows the relationship between two sets of data, $x$ and $\log y$. The regression line of $\log y$ on $x$ is shown and passes through the points (3, 0.7) and (7, 1.3).




(i) Using the given coordinates, find the gradient of the regression line shown.

(ii) Find the regression line of $\log y$ on $x$ in the form $\log y = a + mx$, where $a$ and $m$ are constants to be found.

**(3 marks)**


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

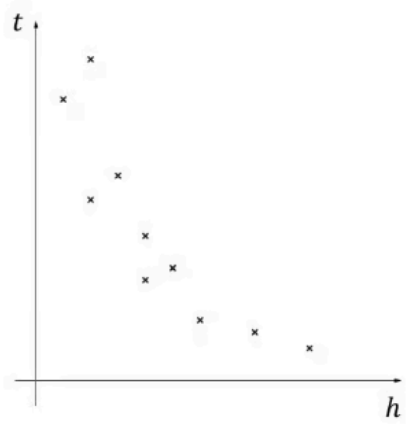
**(c)** (i) By comparing the equations in (a) and (b)(ii) show that $b = 1.778$ to three decimal places.

(ii) Find the value of $p$ to three decimal places.

**(4 marks)**


© 2024 Save My Exams, Ltd.

**8 (a)** The graph below shows the heights, $h$ metres, and the amount of time spent sleeping, $t$ hours, of a group of young giraffes. It is believed the data can be modelled using $t = kh^n$.




By first taking logarithms of both sides, show that $t = kh^n$ can be written as $\log t = \log k + n \log h$ .

**(2 marks)**

**(b)** The data are coded using the changes of variables $x = \log h$ and $y = \log t$. The regression line of $y$ on $x$ is found to be $y = 0.3 - 1.2x$.

(i) Find the values of $x$ and $y$ for a giraffe that is 2.1 metres tall and sleeps for 4.3 hours per day, giving your answers to four decimal places.

(ii) Using the regression line, show that a giraffe of height 3.2 metres would be expected to sleep for approximately half an hour per day.

(iii) State an assumption that was made in order to justify the use of the regression line in part (ii).


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(5 marks)**

**(c)** By first substituting log $h$ for $x$ and log $t$ for $y$ in the equation of the regression line and then by using part (a), show that the relationship between the height of a giraffe and the time it spends sleeping can be modelled by $t = 1.995\ h^{-1.2}$

**(3 marks)**


© 2024 Save My Exams, Ltd.

**9 (a)** The table below shows the daily total sunshine, $x$ hours, and the daily mean total cloud cover, $y$ oktas, for the first 10 days in May in Heathrow in 2015, taken from the large data set.

| $x$ | $a$ | 0.7 | 3.3 | 6.9 | 4.7 | 5.4 | 5.5 | 0.1 | 5.7 | 7.5 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 5 | 6 | 7 | 5 | 6 | 6 | 5 | 7 | 4 | 4 |

Use your knowledge of the large data set to:

(i) explain what is meant by 5 oktas of cloud cover.

(ii) show that $a = 4.4$, given that there were 4 hours and 24 minutes of sunshine recorded in Heathrow on the first day of May 2015.

**(2 marks)**

**(b)** Calculate the product moment correlation coefficient, $r$, for the relationship between $x$ and $y$.

**(2 marks)**

**(c)** (i) State suitable null and alternative hypotheses to test whether there is evidence of negative correlation.

(ii) Use the table of critical values for correlation coefficients in your formula booklet to find the critical value for this test using a 0.5% level of significance.

(iii) Test, at the 0.5% level of significance, whether there is evidence of a negative correlation between daily total sunshine and daily mean total cloud cover in Heathrow during May 2015.


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(4 marks)**


© 2024 Save My Exams, Ltd.

# Medium Questions

**1 (a)** A teacher, Ms Pearman, claims that there is a positive correlation between the number of hours spent studying for a test and the percentage scored on it.

Write down suitable null and alternative hypotheses to test Ms Pearman's claim.

**(1 mark)**

**(b)** Ms Pearman takes a random sample of 25 students and gives them a week to prepare for a test. She records the percentage they score in the test, $s$ %, and the amount of revision they did, $h$ hours.

Ms Pearson calculates the product moment correlation coefficient for these data as $r = 0.874$.

Given that the p-value for the test statistic $r = 0.874$ is 0.0217, test at the 5% level of significance whether Ms Pearman's claim is justified.

**(2 marks)**

**(c)** Ms Pearman decides to use a linear regression model for these data. She calculates the equation for the regression line of $s$ on $h$ to be $s = 21.3 + 5.29h$.

(i) Give an interpretation of the value 21.3 in context.

(ii) Give an interpretation of the value 5.29 in context.

**(2 marks)**


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**2 (a)** The following table shows the number of hours spent learning to drive, $d$, and the number of mistakes made in the driving test, $m$, of ten college students.

| $d$ | 48 | 51 | 51 | 57 | 61 | 68 | 70 | 72 | 73 | 75 |
|---|---|---|---|---|---|---|---|---|---|---|
| $m$ | 19 | 21 | 17 | 12 | 8 | 16 | 7 | 4 | 0 | 1 |

The product moment correlation coefficient for these data is $r = -0.869$. A driving instructor, Dave, believes there is a negative correlation between the number of hours spent learning to drive and the number of mistakes made in the driving test.

(i) Write down suitable null and alternative hypotheses to test Dave's claim.

(ii) Test, at the 1% level of significance, whether Dave's claim is justified, given that the relevant critical value is $-0.7155$.

**(3 marks)**

**(b)** Dave calculates the equation of the regression line of $m$ on $d$ to be $m = 50.7 - 0.642d$.

State, giving a reason, whether or not the correlation coefficient is consistent with the use of a linear regression model.

**(1 mark)**

**(c)** (i) Explain why the linear regression model could be unreliable for predicting the number of mistakes a student would make on their driving test after learning for 30 hours.

(ii) By considering a student who has spent 80 hours learning to drive, give a limitation to the linear regression model.


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(2 marks)**

**3 (a)** The table below shows data from the United States regarding annual per capita chicken consumption (in pounds) and the unemployment rate (% of population) between the years 2005 and 2014:.

| **Year** | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Chicken consumption (pounds)** | 86.4 | 86.9 | 85.5 | 83.8 | 80.0 | 82.8 | 83.3 | 80.8 | 82.3 | 83.8 |
| **Unemployment rate (%)** | 5.08 | 4.62 | 4.62 | 5.78 | 9.25 | 9.63 | 8.95 | 8.07 | 7.38 | 6.17 |

The product moment correlation coefficient for these data is $r = -0.821$.
The critical values for a 10% two-tailed test are ±0.5495.

State what is measured by the product moment correlation coefficient.

**(1 mark)**

**(b)** (i) Write down suitable null and alternative hypotheses for a two-tailed test of the correlation coefficient.

(ii) Show that, at the 10% level of significance, there is evidence that the correlation coefficient is different from zero.

**(3 marks)**

**(c)** A newspaper's headline states:

"Eating chicken is the secret to reducing the unemployment rate in the US!"

Explain whether this headline is fully justified.


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(1 mark)**


© 2024 Save My Exams, Ltd.

**4 (a)** Jessica is researching whether there is a correlation between the productivity of university students and the number of hours sleep they get per night.

Write suitable null and alternative hypotheses to test for linear correlation.

**(1 mark)**

**(b)** Jessica takes a random sample of 25 students, measures their productivity during the day, and records how many hours sleep they had during the previous night. She calculates the product moment correlation coefficient and finds that $r = -0.107$.

The table below gives the critical values, for different significance levels, of the product moment correlation coefficient, r, for a sample of size 25.

| One tail | 10% | 5% | 2.5% | 1% | 0.5% | One tail |
|---|---|---|---|---|---|---|
| Two tail | 20% | 10% | 5% | 2% | 1% | Two tail |
| | 0.2653 | 0.3365 | 0.3961 | 0.4622 | 0.5052 | |

Jessica wishes to test, at the 10% level of significance, whether there is evidence that the correlation coefficient for the population is different from zero.

(i) Find the critical regions for Jessica's test.

(ii) Show that, at the 10% level of significance, there is no evidence of a linear correlation.

**(3 marks)**

**(c)** State, with a reason, whether there could be a relationship between students' hours of sleep and their productivity.

**(1 mark)**


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**5 (a)** During a zombie attack, Richard suspects that the number of flies in the area, $f$, is dependent on the number of zombies, $z$.

Richard is trying to decide whether the correlation is linear or non-linear, so he uses a graphical software package to plot two scatter graphs. **Figure 1** shows the graph of $f$ plotted against $z$, and **Figure 2** show the graph of log $f$ plotted against log $z$.

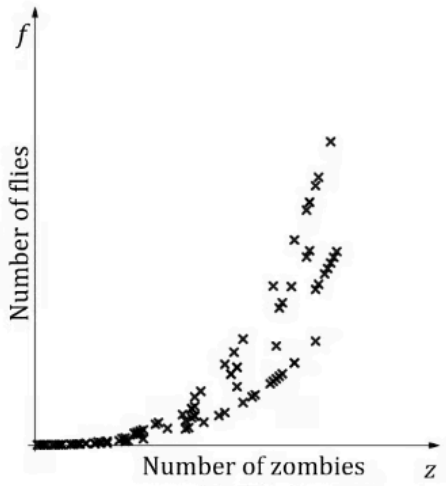


**Figure 1**

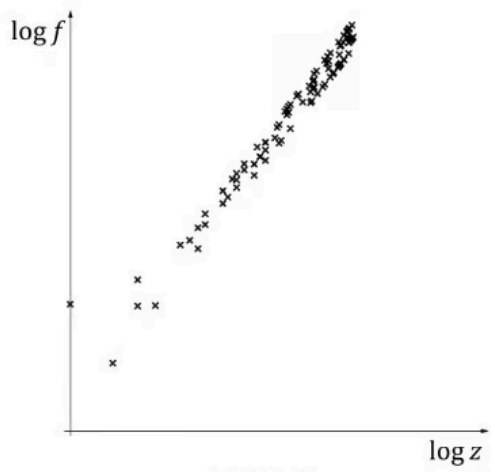


**Figure 2**

Richard calculates the product moment correlation coefficient for each graph. One value is found to be 0.847 while the other is 0.985.

State, with a reason, which PMCC value corresponds with **Figure 2**.

**(2 marks)**

**(b)** Test, using a 5% level of significance, whether there is positive linear correlation in the graph shown in **Figure 2**. State your hypotheses clearly.
You are given that the critical value for this test is 0.1654.

**(3 marks)**


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(c)** State, with a reason, whether the relationship between number of zombies and number of flies is better represented as linear or non-linear.

**(1 mark)**


© 2024 Save My Exams, Ltd.

**6 (a)** Nicole is a Biologist studying the growth of bacteria. She records the number of bacteria on an organism every hour. The table below shows her results for the first eight hours.

| **Hours ($t$)** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **Number of bacteria ($B$)** | 10 | 50 | 170 | 520 | 1730 | 5200 | 17020 | 58140 |

Nicole calculates the product moment correlation coefficient as $r = 0.735$.

Nicole claims that there is a positive linear correlation between the number of hours and the number of bacteria.

Test, at the 1% level of significance, whether Nicole's claim is justified. State your hypotheses clearly.

**(3 marks)**

**(b)** Mariam, Nicole's lab assistant, claims that there is an exponential relationship between the two variables. To test this Mariam calculates the values of $\ln(B)$ for the different values of $t$.

Complete the table, giving your answers to three decimal places.

| $t$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **$\ln(B)$** | 2.303 | 3.912 | 5.136 | 6.254 | 7.456 | 8.556 | | |

**(2 marks)**

**(c)** (i) Calculate the product moment correlation coefficient between $t$ and $\ln(B)$.

(ii) Comment on Mariam's claim that there is an exponential relationship between $B$ and $t$.


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(2 marks)**


© 2024 Save My Exams, Ltd.

**7 (a)** An estate agent, Terry, claims that there is a correlation between the value of a house, $v$ (£1000), and the distance between that house and the nearest nightclub, $d$ (miles).

Terry has a database containing over 100 houses and he takes a random sample of seven houses to investigate his claim. The results are recorded below:

| **d** | 1.8 | 2.1 | 2.5 | 3.7 | 4.9 | 5.2 | 7.2 |
|---|---|---|---|---|---|---|---|
| **v** | 500 | 560 | 330 | 250 | 260 | 180 | 190 |

Calculate the product moment correlation coefficient for this sample.

**(1 mark)**

**(b)** (i) Write down suitable null and alternative hypotheses for a two-tailed test to investigate Terry's claim.

(ii) Test Terry's claim using a 5% level of significance.

**(3 marks)**

**(c)** State, giving a reason, whether the conclusion to the test would be different if a 1% level of significance had been used.

**(1 mark)**

**(d)** Suggest one way in which Terry could improve his investigation.

**(1 mark)**


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

# Hard Questions

**1 (a)** A snack shop owner has noticed that the sale of energy drinks seems to increase later in the school term. He conducts a hypothesis test at the 1% level of significance to see if the sale of the drinks, *d*, increases as the number of days until the school holidays, *h*, decreases.

(i) What type of correlation is the snack shop owner testing for?

(ii) State which of the two variables is the explanatory variable.

**(2 marks)**

**(b)** Over the final thirty days of term the owner keeps a record of the number of sales of energy drinks and, using this data, calculates the product moment correlation coefficient to be $r = -0.4187$.

The table below gives the critical values, for different significance levels, of the product moment correlation coefficient, $r$, for a sample size of 30.

| Level | 10% | 5% | 2.5% | 1% | 0.5% |
|---|---|---|---|---|---|
| $n = 30$ | 0.2407 | 0.3061 | 0.3610 | 0.4226 | 0.4629 |

(i) Write down the critical region for the hypothesis test.

(ii) Stating your hypotheses clearly, test the snack shop owner's suspicion that more energy drinks are sold closer to the school holidays.


© 2024 Save My Exams, Ltd.

**(4 marks)**

**(c)** The snack shop owner calculates the regression line of $d$ on $h$ and uses it to predict the number of energy drinks he will sell on the first day of the new term, when there are still 90 days until the holidays. State two reasons why this is unlikely to give a reliable prediction.

**(2 marks)**


© 2024 Save My Exams, Ltd.

**2 (a)** Adriana is a conservationist researching whether there is any correlation between the population sizes of king cobras, $c$, and their biggest enemy, the Indian grey mongoose, $m$. She collects data on population sizes of both species from a sample of 15 wildlife reserves and calculates the product moment correlation coefficient to be $r = -0.3264$.

The table below gives the critical values, for different significance levels, of the product moment correlation coefficient, $r$, for a sample of size 15.

| One tail | 10% | 5% | 2.5% | 1% | 0.5% | One tail |
|---|---|---|---|---|---|---|
| Two tail | 20% | 10% | 5% | 2% | 1% | Two tail |
| $n = 15$ | 0.3507 | 0.4409 | 0.5140 | 0.5923 | 0.6411 | $n = 15$ |

Conduct a hypothesis test at the 5% level of significance to test if there is linear correlation between $c$ and $m$. Clearly state your hypotheses, critical regions, and conclusion.

**(4 marks)**

**(b)** Adriana concludes that the test indicates that there is no correlation between population sizes of king cobras and the Indian grey mongoose.

Explain why Adriana's conclusion is not fully correct.

**(1 mark)**


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**3 (a)** A biologist is researching a connection between the mass of an animal, $M$ kg, and its expected lifespan, $L$ years. The biologist suggests that there exists a relationship of the form $L=AM^B$, where $A$ and $B$ are constants to be found.

Show that the relationship can be rewritten using logarithms as

$$\log L = \log A + B \log M$$

**(3 marks)**

**(b)** Using data from a wide range of animals, when $y = \log L$ is plotted against $x = \log M$ on a scatter diagram there seems to be a strong positive correlation. When the regression line of $y$ on $x$ is calculated, the equation is found to be $y = 0.18x + 0.98$.

By relating the equation of the regression line to the equation found in (a), or otherwise, find the constants $A$ and $B$ correct to 2 decimal places where appropriate.

**(2 marks)**

**(c)** Hence, predict the lifespan of a horse with a mass of 600 kg to the nearest year.

**(1 mark)**

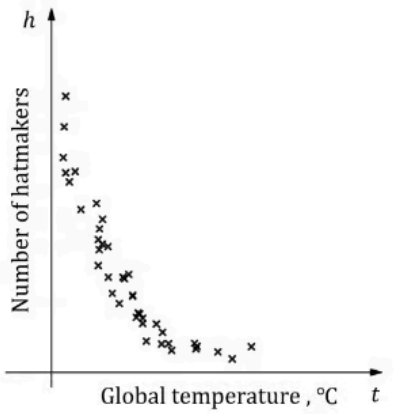
**(d)** The biologist concludes the research by suggesting that one way to increase your lifespan is to increase your mass.

Explain, based on these data, why the biologist may be incorrect.

**(1 mark)**


© 2024 Save My Exams, Ltd.

**4 (a)** M.Hatter has noticed that over the past 50 years there seems to be fewer hatmakers in London. He also knows that global temperatures have been rising over the same time period. He decides to see if there could be any correlation, so he collects data on the number of hatmakers and the global mean temperatures from the past 50 years and records the information in the graph below.




Explain why a model of $h = at + b$ is unlikely to fit these data.

**(1 mark)**

**(b)** Hatter suggests that the equation for $h$ in terms of $t$ can be written in the form $h = ab^t$. He codes the data using $x = t$ and $y = \log h$ and calculates the regression line of $y$ on $x$ to be $y = 1.903 - 1.005x$.

(i) Show that $a = 80.0$ correct to 3 significant figures.

(ii) Find the value of $b$ to 3 significant figures.

(iii) Give an interpretation, in context, of the value of $a$ in your answer to (b)(i).


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(5 marks)**

**(c)** M.Hatter calculates the product moment correlation coefficient between $x$ and $y$ to be $r = -0.952$ and concludes that the rise in mean global temperature is what is causing hatmakers in London to go out of business.

Explain whether M. Hatter's conclusion is fully justified.

**(1 mark)**


© 2024 Save My Exams, Ltd.

**5 (a)** A restaurant owner, Mr Capazio, suspects that there is positive correlation between the number of alcoholic beverages a person has with their meal and the amount of time it takes them to pay their bill at the end of the evening. He decides to conduct a one-tailed hypothesis test at the 5% level of significance to test his theory.

(i) In the context of this question, describe what positive correlation would mean.

(ii) Write down suitable null and alternative hypotheses to test Mr Capazio's theory.

**(2 marks)**

**(b)** The table below shows the number of alcoholic beverages consumed, d, and the amount of time taken to pay the bill, $t$ minutes, for a random sample of 10 visitors to the restaurant.

| Number of drinks, $d$ | 0 | 1 | 3 | 2 | 8 | 4 | 2 | 0 | 3 | 2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Time taken, $t$ minutes | 2.6 | 3.1 | 5.3 | 2.0 | 6.1 | 9.3 | 1.5 | 3.2 | 5.7 | 4.2 |

(i) Find the product moment correlation coefficient for these data.

(ii) Test, at the 5% level of significance, whether there is evidence to suggest Mr Capazio's theory is correct.

**(4 marks)**

**(c)** Mr Capazio calculates the regression line of $t$ on $d$ to be $t = 2.75 + 0.619d$.

(i) Give an interpretation of the values 2.75 and 0.619 in the context of the question.

(ii) A person took 4.5 minutes to pay their bill. Explain why the regression line should not be used to estimate the number of drinks they had had.

**(3 marks)**


© 2024 Save My Exams, Ltd.

**6 (a)** On 21st January 2020, doctors in China started recording and reporting the number of new daily cases of an unknown virus. The doctors record the number of new cases, $c$, of the virus and the number of days, $d$, after 21st January 2020.

| $d$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $c$ | 278 | 48 | 221 | 92 | 277 | $x$ |

| $d$ | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|
| $c$ | 700 | 1700 | 1600 | 1700 | $y$ | 1500 |

The value of the product moment correlation coefficient between the number of days after 21st January 2020 and the number of new cases was calculated as $r = 0.900$.

The table below gives the critical values, for different significance levels, of the product moment correlation coefficient, $r$, for a sample size of 12.

| One tail | 10% | 5% | 2.5% | 1% | 0.5% |
|---|---|---|---|---|---|
| $n = 12$ | 0.3981 | 0.4973 | 0.5760 | 0.6581 | 0.7079 |

(i) Clearly stating suitable null and alternative hypotheses, show that there is evidence of linear correlation between the number of days and the number of new cases at a 1% level of significance.

(ii) Give a reason why a linear regression line is suitable to model the relationship between the number of days and the number of new cases reported each day.

**(4 marks)**


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(b)** The equation for the regression line of $c$ on $d$ is found to be $c = 184.58d - 266.39$.

(i) Use the regression line to estimate the number of new cases on the $6^{th}$ and $11^{th}$ day.

(ii) Explain why the equation for the regression line should not be used to estimate how many new cases there were on $19^{th}$ January 2020.

**(4 marks)**


© 2024 Save My Exams, Ltd.

**7 (a)** Medhi is a meteorologist investigating the weather in Heathrow and claims that there is negative correlation between the daily total rainfall, $f$ mm, and daily total sunshine, $s$ hours.

Write down suitable null and alternative hypotheses to test Medhi's claim.

**(1 mark)**

**(b)** Medhi decides to use the large data set to investigate this claim and forms a sample using all the days in June 2015 relating to Heathrow.

Some values for the daily total rainfall are labelled as "tr". Using your knowledge of the large data set, state the range of values Mehdi could assign to these values.

**(1 mark)**

**(c)** Medhi calculates the product moment correlation coefficient as $r = -0.2659$ using all the days in June.

Test, at the 5% level of significance, whether there is negative correlation between daily total rainfall and daily total sunshine.

**(2 marks)**

**(d)** Medhi uses this data to calculate the equation for the regression line of $f$ on $s$. He plans to use the regression line to predict the amount of rainfall there will be in Heathrow during a day in December.

Give two reasons why this is unlikely to produce a reliable estimate.

**(2 marks)**


© 2024 Save My Exams, Ltd.

# Very Hard Questions

**1 (a)** A doctor is collecting data on how a certain illness affects the weight of a person. Let $w$ be the number of weeks that a patient had the illness and $d$ *kg* be the amount of weight that the patient lost whilst they were ill. The doctor suspects that $d$ and $w$ have a relationship of the form $d = aw^b$, where $a$ and $b$ are constants to be found.

After plotting $\log d = y$ against $\log w = x$, the doctor found there to be a strong correlation and the equation of the regression line of $y$ on $x$ was $y = 1.47x - 0.11$.

Explain why the relationship between $x$ and $y$ must have shown positive correlation.

**(1 mark)**

**(b)** By using the equation of the regression line of $y$ on $x$, or otherwise, find the values of the constants $a$ and $b$ correct to 3 significant figures where appropriate.

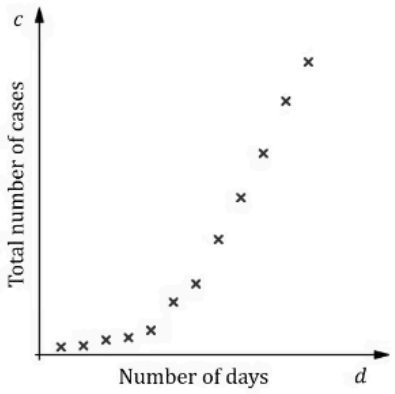
**(4 marks)**

**(c)** Stating any assumptions you make, find the weight loss expected of a patient who has been sick for 20 days to the nearest whole kilogram.

**(3 marks)**


© 2024 Save My Exams, Ltd.

**2 (a)** Scientists in Wuhan, China, started tracking the total number of cases of the CoViD-19 virus in January 2020. The graph below shows the number of days, $d$, after the first reported case, and the total number of cases, $c$, of the virus for a period of 12 days.




(i) Give a reason why the scientists should not use a regression line to model the relationship between the number of days and the total number of cases.

(ii) After two days the scientists tried to model the relationship using an exponential model of the form $c = kb^d$. Given that after 1 day there were 278 cases and after 2 days there were 326 cases, calculate the values of $k$ and $b$.

(iii) After 11 days there were 9700 reported cases in China. Comment on the suitability of this model.

**(5 marks)**


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(b)** Another group of scientists code the data using $x = d$ and $y = \log c$. The regression line of $y$ on $x$ was found to be $y = 2.2476 + 0.1606x$.

(i) Using the regression line of $y$ on $x$, find an equation for $c$ in terms of $d$ in the form $c = ap^d$. State the values of $a$ and $p$ to 4 significant figures.

(ii) Explain what the value of $p$ represents in your answer to (b)(i).

(iii) One of the scientists used this model and predicted after three months there would have been over $4.89 \times 10^{16}$ cases. Give two reasons to explain why this prediction is unlikely.

**(5 marks)**


© 2024 Save My Exams, Ltd.

**3 (a)** Rory is studying the relationship between two variables $x$ and $y$. He believes they could be modelled by the equation $y = ax^m$ where $a$ and $m$ are constants. He codes his data and plots a scatter graph of $X = \log x$ against $Y = \log y$. Rory draws, by eye, a line of best fit between $X$ and $Y$ which passes through the points (2, 2.68) and (5, 3.10).

Using Rory's line of best fit, find the values of the constants $a$ and $m$ in his model. Give your answers correct to 3 significant figures where appropriate.

**(5 marks)**

**(b)** Rory wants to conduct a test, using a 1% level of significance, to test whether there is a positive linear correlation between $X$ and $Y$. The value of the product moment correlation coefficient for $X$ and $Y$ is calculated to be $r = 0.5047$ for Rory's data.

(i) Given that the critical value for this test is 0.6581, write down the size of Rory's sample.

(ii) Stating your hypotheses clearly, complete Rory's hypothesis test.

(iii) Comment on the suitability of Rory's equation $y = a\,x^m$.

**(5 marks)**


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(c)** Rory later discovers the relationship between $x$ and $y$ is better suited to the equation model $y = kb^x$. From his raw data he calculates the regression line of log $y$ on $x$ to be $\log y = 2.56 + 1.12x$.

Find the value of the constants $k$ and $b$ in Rory's new model, giving your answers to 3 significant figures.

**(3 marks)**


© 2024 Save My Exams, Ltd.

**4 (a)** Alfie lives near a river and from observing the wildlife he believes that there is a relationship between the number of otters and the number of frogs. He decides to investigate this further and gathers data on frog and otter populations from six wildlife centres around the UK. Alfie records the data in the table below.

| **Frogs, $f$** | 8500 | 2000 | 6000 | 5000 | 3000 | 1500 |
|---|---|---|---|---|---|---|
| **Otters, t** | 100 | 300 | 110 | 130 | 120 | 500 |

The product moment correlation coefficient for these data is $r = -0.7401$.

(i) Stating your hypotheses clearly, test at the 10% level of significance, whether there is linear correlation between the number of frogs and the number of otters in the wildlife.

(ii) The figures in the table are given to the nearest hundred. Considering this, comment on the validity of the conclusion to the test.

**(5 marks)**

**(b)** Alfie codes the data using logarithms and records the results in the table below.

| **log $f$** | 3.329 | 3.301 | 3.778 | 3.699 | 3.477 | 3.176 |
|---|---|---|---|---|---|---|
| **log $t$** | 2.000 | 2.477 | 2.041 | 2.114 | 2.079 | 2.699 |

(i) Calculate the product moment correlation coefficient between log $f$ and log $t$ using the data from the table above.

(ii) Show that, at the 10% level of significance, there is not enough evidence to suggest there is a linear correlation between log $f$ and log $t$.


© 2024 Save My Exams, Ltd. | Get more and ace your exams at savemyexams.com

**(3 marks)**

**(c)** Alife concludes that the evidence from the test suggests that there is not a non-linear relationship between the number of frogs and the number of otters.

Explain why Alfie could be incorrect by describing a type of relationship that might exist between the number of frogs and the number of otters

**(1 mark)**

**5** A researcher has been collecting data within a particular city on the number of sleeveless T-shirts sold per week, $T$, and the number of new gym memberships per week, $G$. The data is shown in the table below along with the values of log $T$ and log $G$.

| $T$ | 119 | 54 | 92 | 25 | 442 | 340 | 9 | 261 |
|---|---|---|---|---|---|---|---|---|
| $G$ | 50 | 15 | 25 | 12 | 129 | 22 | 8 | 21 |
| log $T$ | 2.0755 | 1.7324 | 1.9638 | 1.3979 | 2.6454 | 2.5315 | 0.9542 | 2.4166 |
| log $G$ | 1.6990 | 1.1761 | 1.3979 | 1.0792 | 2.1106 | 1.3424 | 0.9031 | 1.3222 |

The researcher suspects that $T$ and $G$ are related in one of two ways:

$$T = aG^m \text{ or } T = bp^G$$

where $a$, $b$, $m$, and $p$ are constants.

By calculating the product moment correlation coefficient for two different pairs of rows in the table, decide which of the two models better represents the relationship between $T$ and $G$.

**(5 marks)**


© 2024 Save My Exams, Ltd.

**6 (a)** Charlie is interested to find out if there is positive correlation between the number of letters in someone's name, $l$, and the time, $t$ rounded to the nearest five seconds, it takes her six-year-old sister to correctly guess the spelling of the name. She decides to test this by looking at a random sample of different names and timing how long it takes her sister to guess their spelling.

| **Letters,** $l$ | 3 | 3 | 4 | 4 | 4 | 5 | 5 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Time,** $t$ | 0 | 5 | 5 | 10 | 15 | 5 | 15 | 25 | 60 | 80 |
| **Frequency** | 2 | 16 | 4 | $x$ | 17 | 3 | 29 | 17 | 4 | 1 |

Charlie carries out the one-tailed test at the 1% level of significance and finds that the critical value for this test is 0.2324.

Find the total number of names in Charlie's sample and hence, find the value of $x$.

**(2 marks)**

**(b)** Stating your hypotheses clearly, test, at the 1% level of significance, if there is evidence of positive linear correlation between the number of letters in someone's name and the time it takes Charlie's sister to guess the correct spelling.

**(3 marks)**

**(c)** Charlie calculates the equation of the regression line of $t$ on $l$ to be t = 1 1.2$l$ −33.4.

(i) Give an interpretation of the value of the gradient of this regression line.

(ii) Explain why Charlie should not use this equation to estimate the number of letters that are in someone's name if it took her sister 70 seconds to guess the spelling.

**(3 marks)**


© 2024 Save My Exams, Ltd.

**7 (a)** When brewing beer the temperature that the beer is stored at during fermentation, T ℃, changes the alcohol content, $A$ %, at the end of the fermentation process. A group of brewers collect data on $T$ and $A$ for their casks of beer. They suspect the data follows a model of the form $A = bp^T$ where $b$ and $p$ are unknown constants. They plot the regression line of $y = \ln A$ on $x = T$ and find that the line has a gradient of 0.0392 and passes through the point (0, 0.811).

Using the line of regression, calculate the values of $b$ and $p$, giving your answers correct to 3 significant figures.

**(5 marks)**

**(b)** In the data collected by the brewers, the range of values for $T$ was 15 and the range of values for $A$ was 4. You are given that the minimum alcohol content occurred when the temperature was at its minimum and the maximum alcohol content occurred when the temperature was at its maximum.

Find estimates for the minimum values of $T$ and $A$ to 2 significant figures.

**(7 marks)**


© 2024 Save My Exams, Ltd.

**(c)** Hence explain why it would not be appropriate to use the model to predict the alcohol content of beer when the temperature during fermentation is 50 ℃.

**(1 mark)**


© 2024 Save My Exams, Ltd.