STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

Correlation

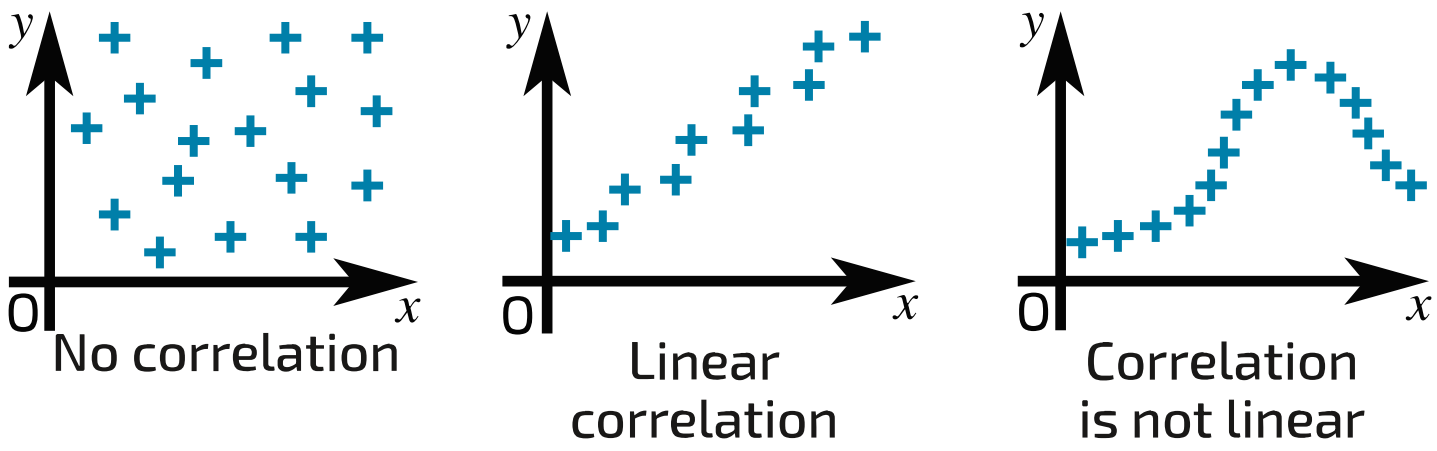
A-level Maths Topic Summaries - Statistics

Subject & topics: Maths | Statistics | Data Analysis **Stage & difficulty:** A Level P3

Fill in the blanks to complete these notes about correlation.

Part A

Correlation



A [ ] or **scatter diagram** uses paired data values as coordinates for a graph. One variable is assigned to the x-axis, and the other to the y-axis. If a pattern can be seen in the plotted data then the variables are [ ]. When points cluster along a straight line the correlation is [ ].

Correlation between two variables [ ] imply that one of the variables is responsible for causing a change in the other variable. This **may or may not** be the case. The correlation may reflect the fact that both variables are dependent on a third variable. For example, the height and weight of a growing plant both depend on time. This is summed up by the sentence **"Correlation does not imply** [ ]**."**

Items:

bivariate plot causation correlated does not linear

Part B

Linear correlation and the pmcc

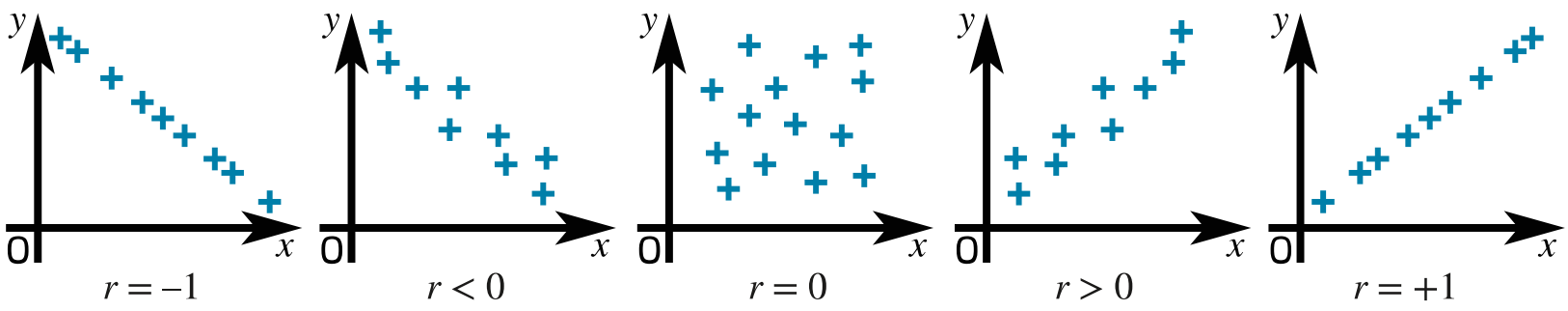


Figure 1: Illustrating the different values of r.

The **pmcc** (product moment correlation coefficient) is used to describe the degree and type of [ ] correlation between two variables. In other words, it measures how closely the points on a scatter diagram lie along a straight line.

The letter [ ] is used for the pmcc. The pmcc takes a value between -1 and $+1$.

- If $r < 0$, there is [ ] correlation
- If $r = 0$ then there is [ ] correlation
- If $r > 0$ there is [ ] correlation
- If $r = -1$ or $r = +1$ there is [ ] correlation between the variables. The points on the diagram all lie exactly along a straight line.

Items:

linear | negative | no linear | perfect linear | positive | r

Part C

Hypothesis tests for linear correlation

When we sample a population, an important question is whether the value of r that we calculate from the sample is sufficiently large to conclude that there is evidence of a linear relationship between the two variables in the population as a whole. We use a hypothesis test to decide this.

We use ρ for the correlation coefficient in the population. Our null hypothesis is always that there is [ ] in the population, $\mathrm{H}_0 : \rho = 0$.

The alternative hypothesis can take one of three forms. We can test for linear correlation of any kind [ ], which is a two-tailed test. Or, we can test for positive linear correlation [ ] or negative linear correlation [ ], which are one-tailed tests.

To carry out the test we need to look up the critical value of the correlation coefficient in a table. The critical value depends on the sample size n, the significance level, and whether the test is one- or two-tailed. We then compare the value of r from the sample to the critical value and state the conclusion.

- If $|r| < r_{\text{crit}}$, we [ ] the null hypothesis. There [ ] significant evidence for the alternative hypothesis.
- If $|r| > r_{\text{crit}}$, we [ ] the null hypothesis. There [ ] significant evidence for the alternative hypothesis.

Items:

do not reject | is | is not | no correlation | reject | $(\mathrm{H}_1 : \rho < 0)$ | $(\mathrm{H}_1 : \rho \neq 0)$ | $(\mathrm{H}_1 : \rho > 0)$

Created for isaacscience.org by Jonathan Waugh

isaac You work it out.

STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

Linear regression 3.1

Subject & topics: Maths | Statistics | Hypothesis Tests **Stage & difficulty:** A Level P2

An experiment was carried out to find the acceleration of a ball rolling down a straight ramp. Assuming that the ball starts from rest and that the acceleration is constant, the relationship between the time t it takes to roll a distance s down the ramp is such that

$$s = \frac{1}{2}at^2$$

where a is the acceleration.

Two students A and B carry out the experiment one after the other using the same apparatus but making their own measurements of s and t. They decide to measure the time it takes for the ball to roll a fixed distance i.e. the distance is the independent variable and the time is the dependent variable.

Part A

Rearrange the equation

Rearrange the expression

$$s = \frac{1}{2}at^2$$

to obtain an equation for t (i.e. make t the subject of the equation).

The following symbols may be useful: `a, s, t`

Part B

A's data

A graph of time against the square root of distance using A's data is shown in **Figure 1**; the time t is in s and the square root of the distance $\sqrt{s}$ is in $\mathrm{cm}^{\frac{1}{2}}$. The uncertainties in the values are smaller than the sizes of the crosses. The line of best fit, assuming that it goes through the point $(0, 0)$, is shown; it has the following parameters,

$$t = 0.4137\sqrt{s} \qquad r = 0.9998 \qquad r^2 = 0.9996$$

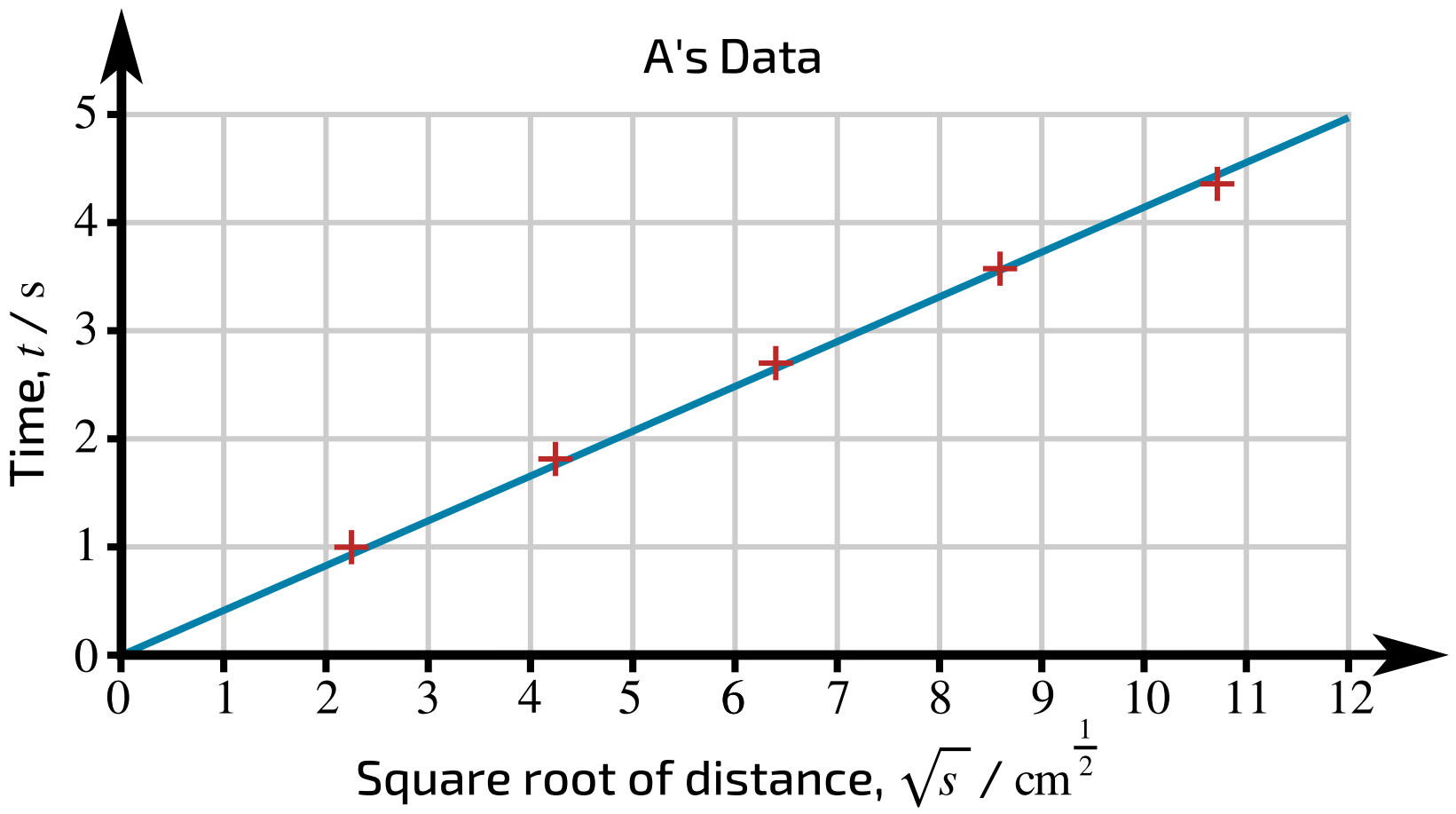


Figure 1: A graph of t against $\sqrt{s}$ using A's data, showing the line of best fit.

Deduce the value of the acceleration down the ramp.

Part C

B's data

A graph of time against the square root of distance using B's data is shown in **Figure 2**; the time t is in s and the square root of the distance $\sqrt{s}$ is in $\mathrm{cm}^{\frac{1}{2}}$. The uncertainties in the values are smaller than the sizes of the crosses. The line of best fit, assuming that it goes through the point $(0, 0)$ and using all five points, is shown. It is clear from looking at the graph that although the r value is very high, there is an obvious outlier; on discussing the results with Student A, Student B concludes that they have either measured or recorded the distance incorrectly in this case. The line is fitted again omitting the outlier.

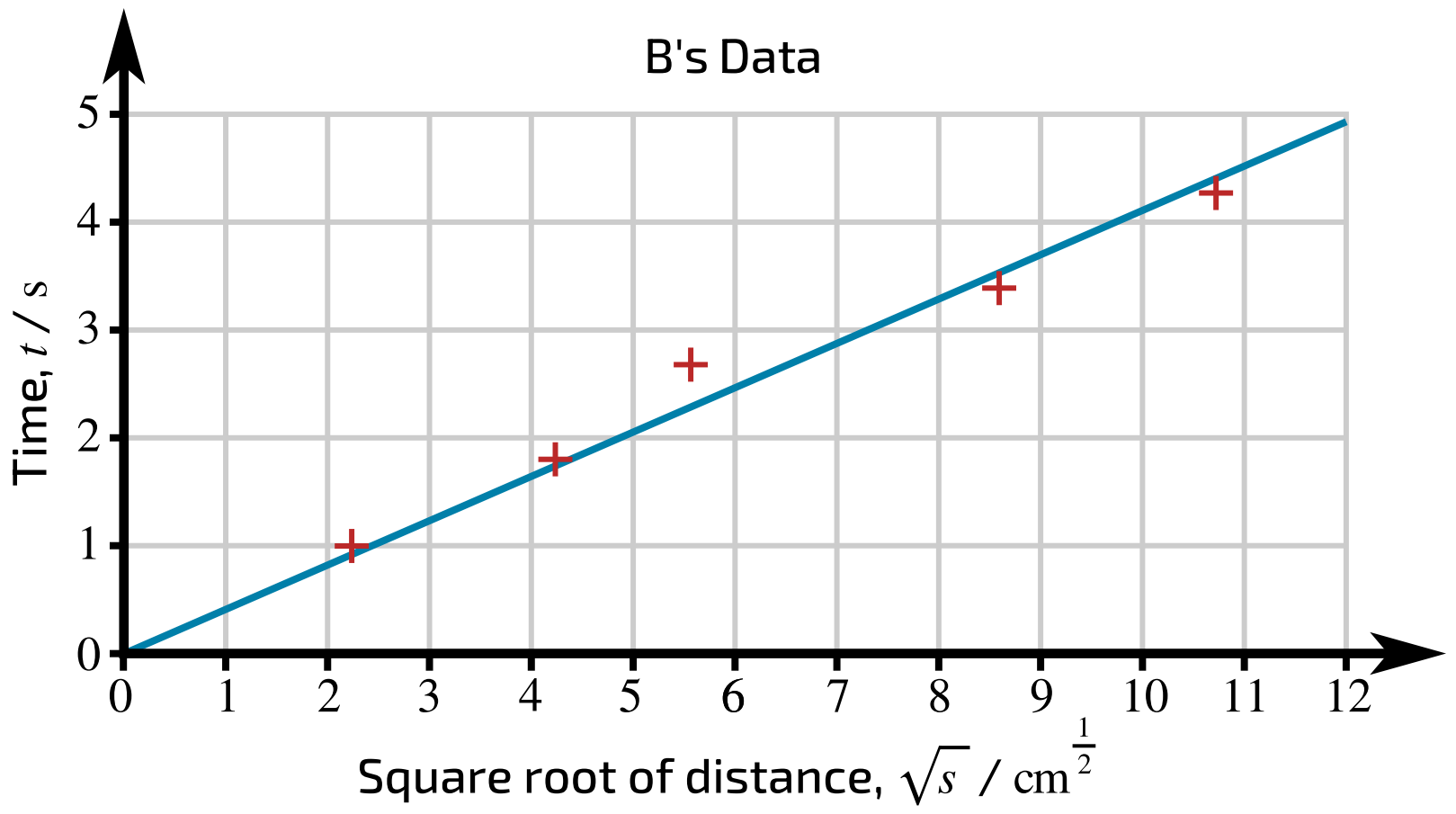


Figure 2: A graph of t against $\sqrt{s}$ using B's data, showing the line of best fit.

The lines of best fit have the following parameters:

$$t = 0.4109\sqrt{s} \qquad r = 0.9976 \qquad r^2 = 0.9951$$

$$t = 0.4005\sqrt{s} \qquad r = 0.9997 \qquad r^2 = 0.9993.$$

Decide which of the two you think is the line of best fit when the outlier is omitted and give its r value.

Part D

Acceleration using B's data

Find the value of the acceleration down the ramp using the more appropriate set of B's data.

Part E

Time to roll 150 cm

Use A's data to estimate the time it will take the ball to roll $1.5\,\mathrm{m}$ down the ramp giving your answer to 3 sf.

Part F

Difference in time estimates

The line of best fit to the data can be used to estimate the time it will take the ball to roll any given distance down the ramp. Find the percentage difference between the estimates obtained by A and B, giving your answer to 1 sf.

Created for isaacphysics.org by Julia Riley

Question deck:

STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

isaac You work it out.

STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

Correlation Hypothesis Testing 1

Subject & topics: Maths | Statistics | Hypothesis Tests **Stage & difficulty:** A Level P3

In each part, carry out a hypothesis test for the requested type of correlation at the stated significance level.

Part A
Positive correlation

A sample of size $n = 17$ has a correlation coefficient of $r = 0.601$. Test at the 1% significance level whether the population from which the sample was taken has positive correlation, then fill in the blanks below.

The null hypothesis is that the population has no correlation. The alternative hypothesis is that the population exhibits positive correlation.

$\mathrm{H}_0 : \rho = 0$ [ ]

For a one-tailed test, the critical value of the correlation coefficient for a sample of size 17 is [ ] at the 1% significance level.

The correlation coefficient for the sample is 0.601. This is [ ] than the critical value. Hence, we [ ] the null hypothesis. There is evidence that the population exhibits positive correlation.

Items:

0.5577 0.5742 0.6055 $\mathrm{H}_1 : \rho < 0$ $\mathrm{H}_1 : \rho \neq 0$ $\mathrm{H}_1 : \rho > 0$ reject do not reject smaller equal to larger

Part B

Negative correlation

A researcher believes that average speeds for athletes running the $400\,\mathrm{m}$ at a particular track are lower if there has been a larger amount of rainfall earlier in the day. The researcher times one particular athlete at the same time every day on ten different autumn days. They record the depth of rainfall (in mm) before each run, and work out the athlete's average speed. The researcher calculates that the correlation coefficient is -0.713. Test at the 5% significance level whether an athlete's average speed and the amount of rainfall are indeed negatively correlated at this track.

Available items

1. The null hypothesis is that there is no correlation. The alternative hypothesis is there is negative correlation.

1. The null hypothesis is that there is negative correlation. The alternative hypothesis is that there is no correlation.

2. $\mathrm{H}_0 : \rho = 0$ $\mathrm{H}_1 : \rho > 0$

2. $\mathrm{H}_0 : \rho = 0$ $\mathrm{H}_1 : \rho < 0$

3. For a one-tailed test, the critical value of the correlation coefficient for a sample of size 5 is 0.8054 at the 10% significance level.

3. For a one-tailed test, the critical value of the correlation coefficient for a sample of size 10 is 0.5494 at the 5% significance level.

4. The correlation coefficient for the sample is -0.713. This is negative, and has a magnitude greater than the critical value ($|-0.713| > 0.5494$).

4. The correlation coefficient for the sample is -0.713, and $-0.713 < 0.5494$.

5. Hence, we do not reject the null hypothesis. There is not significant evidence for negative correlation between an athlete's average speed and the amount of rainfall.

5. Hence, we reject the null hypothesis. There is evidence that an athlete's average speed and the amount of rainfall have negative correlation.

Part C

Any (linear) correlation

An author wonders whether the amount of time their cat sits next to them is correlated with the number of words they write during the day. Over fifty-three days, the author records the number of words they write and for how long the cat sits nearby, and finds $r = 0.3300$. Test the data at the 1% significance level.

Choose from the options below to construct a complete hypothesis test.

This question is looking for correlation in either direction. A two-tailed test is needed.

- [ ] $\mathrm{H}_0 : \rho = 0 \qquad \mathrm{H}_1 : \rho \neq 0$

This question is looking for positive correlation. A one-tailed test is needed.

- [ ] $\mathrm{H}_0 : \rho = 0 \qquad \mathrm{H}_1 : \rho > 0$

- [ ] For a one-tailed test, the critical value of the correlation coefficient for a sample of size 53 is 0.3188 at the 1% significance level.

- [ ] For a two-tailed test, the critical value of the correlation coefficient for a sample of size 53 is 0.3509 at the 1% significance level.

- [ ] The correlation coefficient for the sample is 0.3300, and $0.3300 < 0.3509$.

- [ ] The correlation coefficient for the sample is 0.3300, and $0.3300 > 0.3188$.

- [ ] Hence, we do not reject the null hypothesis. There is not significant evidence that the number of words the author writes is correlated with the amount of time their cat sits near them.

- [ ] Hence, we reject the null hypothesis. There is evidence that the number of words the author writes is correlated with the amount of time their cat sits near them.

Created for isaacphysics.org by Jonathan Waugh

Question deck:

STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

isaac You work it out.

STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

Correlation Hypothesis Testing 2

Subject & topics: Maths | Statistics | Hypothesis Tests **Stage & difficulty:** A Level P3

A town planner believes that on summer weekday afternoons the amount of traffic into the centre of their town is higher when the temperature is higher. They want to test this hypothesis at the 1% significance level.

Every weekday (Monday to Friday) for six weeks they monitor the traffic on the main roads into town, and record the mean afternoon temperature, and they find that the correlation coefficient is -0.4517.

Part A

Initial conclusion

Without doing any calculations, which of these statements can the town planner make? Choose all that apply.

- [ ] The correlation coefficient is negative, so there is no evidence that the amount of traffic is positively correlated with temperature.
- [ ] The correlation coefficient is negative. There may be a negative correlation between the amount of traffic and afternoon temperature.
- [ ] The correlation coefficient of their sample is negative, so there is a negative correlation between the amount of traffic on summer afternoons and temperature.
- [ ] There is no evidence that the amount of traffic and afternoon temperature are correlated.
- [ ] The correlation coefficient is negative, so there is definitely no positive correlation between the amount of traffic and temperature.

Part B

Choosing a hypothesis test

Using the given data, which of the following hypothesis tests would it be most useful for the town planner to carry out?

- A hypothesis test to see if the amount of traffic and afternoon temperature are negatively correlated at the 1% significance level.
- A hypothesis test to see if the amount of traffic and afternoon temperature are negatively correlated at the 20% significance level.
- A hypothesis test to see if the amount of traffic and afternoon temperature are negatively correlated at the 50% significance level.

Part C

Null and alternative hypotheses

The town planner carries out the most useful test listed in part B.

Drag and drop symbols into the spaces below to state the null and alternative hypotheses for this test, where ρ represents the population correlation coefficient and r represents the sample correlation coefficient.

H_0: ☐☐☐

H_1: ☐☐☐

Items:

ρ r $<$ $=$ $>$ 0 1

Part D

Carrying out the test

Carry out the hypothesis test, and make a conclusion. Then fill in the blanks below.

The critical value of the correlation coefficient is [ ].

Comparing the town planner's value to the critical value gives [ ].

Therefore, [ ] the null hypothesis. There [ ] significant evidence that there is a [ ] correlation between the amount of traffic in summer and afternoon temperature.

Items:

0.4226	0.8822	0.9343	$\lvert -0.4517\rvert > 0.4226$	$\lvert -0.4517\rvert < 0.9343$	$\lvert -0.4517\rvert < 0.8822$	reject
do not reject	is	may be	is not	negative	positive	no

Created for isaacphysics.org by Jonathan Waugh

Question deck:

STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

Linear regression 3.3

Subject & topics: Maths | Statistics | Hypothesis Tests **Stage & difficulty:** A Level C1

A graph of Hubble's original data relating the recession velocity v of a galaxy to its distance D from us is shown in **Figure 1**; the velocity v is in $\mathrm{km\ s^{-1}}$ and the distance D is in Mpc. (Distances in astronomy are often measured in parsecs (abbreviated to pc), where $1\,\mathrm{pc} = 3.26$ light-years $= 3.09 \times 10^{16}\,\mathrm{m}$ and $1\,\mathrm{Mpc} = 10^6\,\mathrm{pc}$.)

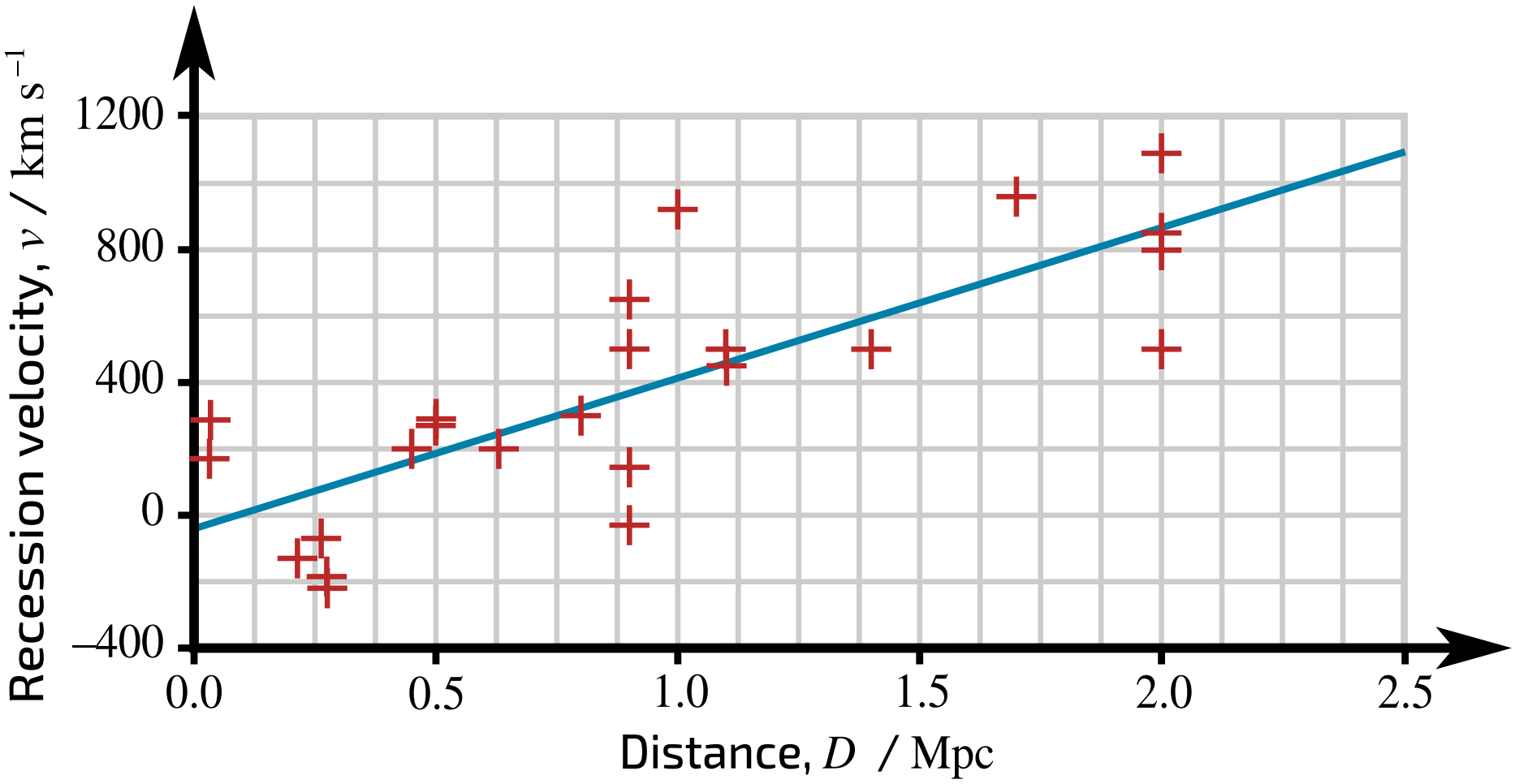


Figure 1: A graph of Hubble's original data relating the recession velocity v of a galaxy to its distance D from us. The regression line of best fit is shown.

The equation describing the best fit to the data is of the form $v = a + bD$ and has the following parameters

$$a = -40.8 \quad b = 454.2 \quad r = 0.790 \quad r^2 = 0.623.$$

Part A

The units of a

What are the units of a?

- $\mathrm{Mpc/km\,s^{-1}}$
- $\mathrm{Mpc\,s\,km^{-1}}$
- $\mathrm{s\,km^{-1}}$
- $\mathrm{km\,s^{-1}}$
- $\mathrm{km\,s^{-1}\,Mpc^{-1}}$
- $\mathrm{km\,s^{-1}/Mpc}$

Part B

The units of b

What are the units of b? (The quantity b is called the Hubble constant and is usually written H_0).

- $\mathrm{km\,s^{-1}/Mpc}$
- $\mathrm{Mpc\,km\,s^{-1}}$
- $\mathrm{Mpc\,s/km}$
- $\mathrm{Mpc/km\,s^{-1}}$
- Mpc
- $\mathrm{km\,s^{-1}\,Mpc^{-1}}$

Part C

Recession velocity

Using the best fit equation above estimate the recession velocity of a galaxy at a distance of 6.0×10^6 light years; give your answer to 2 sf.

Part D

The age of the Universe using the original data

Nowadays the value of the Hubble constant is known to be close to 70 in the same units as b. (It is significantly smaller than that originally determined by Hubble because of a calibration error in Hubble's original data.) The equation describing the relationship between v and D in the same units as above is therefore

$$v = 70D.$$

It is straightforward to show that the age of the Universe is given by $\frac{1}{H_0}$, where H_0 is the Hubble constant.

Find the age of the Universe using the value of b estimated from Hubble's original data above. Give your answer in years and to 2 sf.

Part E

The age of the Universe using current data

Find the age of the Universe using the current value of $H_0 = 70$ (in the same units as b). Give your answer in years and to 2 sf.

Created for isaacphysics.org by Julia Riley

Question deck:

STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

isaac You work it out.

STEM SMART Single Maths 40 - Correlation & Hypothesis Testing

Linear regression 3.2

Subject & topics: Maths | Statistics | Hypothesis Tests **Stage & difficulty:** A Level C1

An experiment is carried out to measure the resistance R of a semiconductor as a function of absolute temperature T. Theory suggests that above a certain temperature

$$R = R_0 \, \mathrm{e}^{\frac{b}{T}}$$

where R_0 and b are constants.

Part A
Rearrange the equation

By taking the natural logarithms of both sides of the equation show that it can be written

$$\ln R = a + f(T)$$

where a is a constant and $f(T)$ is a function of T. Find expressions for a and $f(T)$.

Find an expression for a.

The following symbols may be useful: `R_0, T, a, b, e, f(T)`

Find $f(T)$.

The following symbols may be useful: `R_0, T, a, b, e, f(T)`

Part B

Lines of best fit

In **Figure 1**, $\ln R$, where R is in $\mathrm{k\Omega}$, is plotted as a function of $\frac{1}{T}$, where $\frac{1}{T}$ is in $10^{-3}\,\mathrm{K}^{-1}$. Thus if $R = 1000\,\Omega$, then $\ln R = \ln(1\,\mathrm{k\Omega}) = 0$ and, if $T = 500\,\mathrm{K}$, then $\frac{1}{T} = 0.002\,\mathrm{K}^{-1} = 2 \times 10^{-3}\,\mathrm{K}^{-1}$.

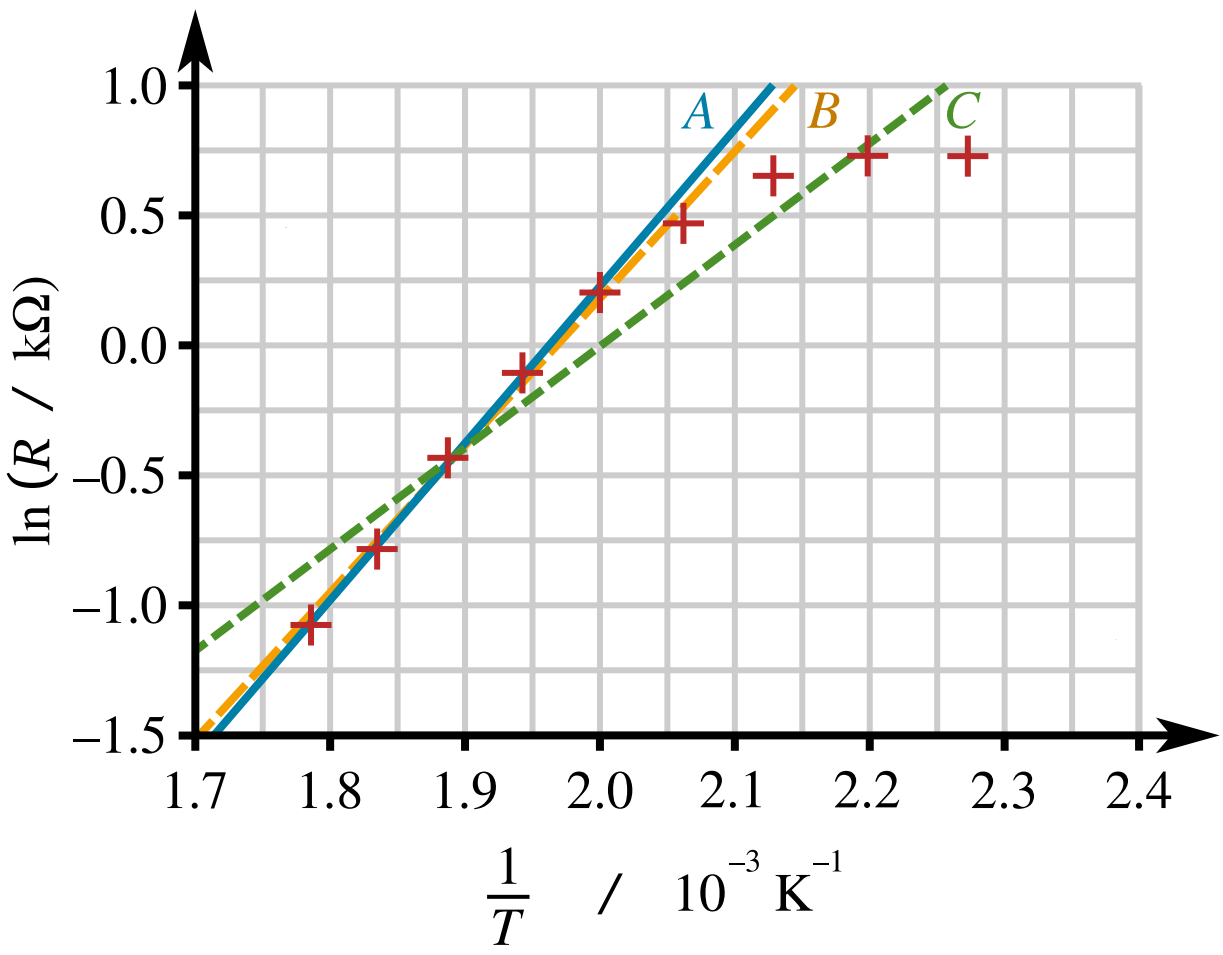


Figure 1: A plot of $\ln R$ against $\frac{1}{T}$; three lines of best fit, A (blue, solid), B (yellow, long-dashed) and C (green, short-dashed) fitted to different ranges of the data, are shown.

Three lines of best fit A, B and C are fitted to different ranges of the data as shown in **Figure 1**.

The parameters of the three fitted lines are:

$$\begin{array}{rllll} \mathrm{I}: & a = -11.197 & b = 5.686 & r = 0.997 & r^2 = 0.994 \\ \mathrm{II}: & a = -11.857 & b = 6.042 & r = 0.999 & r^2 = 0.998 \\ \mathrm{III}: & a = -7.816 & b = 3.906 & r = 0.955 & r^2 = 0.913 \end{array}$$

where a and b are as defined in Part A and the initial equation, and b has units of $10^3\,\mathrm{K}$.

Match the lines to the parameters.

I corresponds to line ☐.

II corresponds to line ☐.

III corresponds to line ☐.

Items:

A B C

Part C

Deductions from the graphs

Theory suggests that above a certain temperature

$$R = R_0 \, \mathrm{e}^{\frac{b}{T}}$$

where R_0 and b are constants.

Using the information from the graphs in part B, suggest, to 1 sf, the temperature above which the theory is valid.

Part D

Estimate the energy gap

According to the theory the energy gap between the insulating and the conducting energy bands in a semiconductor is $E_g = 2kb$, where k is the Boltzmann constant ($k = 1.4 \times 10^{-23} \, \mathrm{J \, K^{-1}}$). Select the line of best fit from Part B which best fits the theory and deduce the value of b; hence estimate E_g.

Part E

Resistance when $T = 520 \, \mathrm{K}$

Use the equation of the line of best fit from Part B to deduce the resistance at $520 \, \mathrm{K}$.

Part F

Resistance when $T = 450\,\mathrm{K}$

By looking at the graph in Part B, deduce the resistance when $T = 450\,\mathrm{K}$ giving your answer to 1 significant figure.

Created for isaacphysics.org by Julia Riley