Model Solutions

Additional Assessment Materials

Summer 2021

Pearson Edexcel GCE in Mathematics

9MA0 (Applied) (Public release version)

Resource Set 1: Topic 5

Statistical hypothesis testing

Pearson: helping people progress, everywhere

Pearson aspires to be the world's leading learning company. Our aim is to help everyone progress in their lives through education. We believe in every kind of learning, for all kinds of people, wherever they are in the world. We've been involved in education for over 150 years, and by working across 70 countries, in 100 languages, we have built an international reputation for our commitment to high standards and raising achievement through innovation in education. Find out more about how we can help you and your students at: www.pearson.com/uk

Additional Assessment Materials, Summer 2021

General guidance to Additional Assessment Materials for use in 2021

Context

- Additional Assessment Materials are being produced for GCSE, AS and A levels (with the exception of Art and Design).
- The Additional Assessment Materials presented in this booklet are an optional part of the range of evidence teachers may use when deciding on a candidate's grade.
- 2021 Additional Assessment Materials have been drawn from previous examination materials, namely past papers.
- Additional Assessment Materials have come from past papers both published (those materials available publicly) and unpublished (those currently under padlock to our centres) presented in a different format to allow teachers to adapt them for use with candidate.

Purpose

- The purpose of this resource to provide qualification-specific sets/groups of questions covering the knowledge, skills and understanding relevant to this Pearson qualification.
- This document should be used in conjunction with the mapping guidance which will map content and/or skills covered within each set of questions.
- These materials are only intended to support the summer 2021 series.

1. Tessa owns a small clothes shop in a seaside town. She records the weekly sales figures, £ w, and the average weekly temperature, t °C, for 8 weeks during the summer.

The product moment correlation coefficient for these data is −0.915.

(a) Stating your hypotheses clearly and using a 5% level of significance, test whether or not the correlation between sales figures and average weekly temperature is negative.

(3)

let r denote the correlation coefficient, then our test hypothesis will be one sided since we only want to see if the correlation between the two variables is negative. So let; $H_0 : r = 0$ v.s. $H_1 : r < 0$

we then use critical value table, noting that $n = 8$ since we have 8 weeks of data and $\alpha = 0.05$, which means that the critical value is -0.6215.

we have that $-0.915 < -0.6215$ holds true which means that the negative critical value is greater than the correlation coefficent. This means that r is significant => we reject H_0 and conclude that $r < 0$, i.e there is a significant correlation between weekly sales and temperature.

(b) Suggest a possible reason for this correlation.

(1)

As the temperature increases, sales go down. This could be down to the types of clothes Tessa sells (i.e. the correlation suggests she may sell winter clothes hence less are sold when its warmer), they may also be more likely to go outdoors.

Tessa suggests that a linear regression model could be used to model these data.

(c) State, giving a reason, whether or not the correlation coefficient is consistent with Tessa's suggestion.

(1)

The correlation coefficient is close to −1, which means that a linear regression model could work well.

(d) State, giving a reason, which variable would be the explanatory variable.

(1)

Explanatory variable is the independent variable, and will always go on the x-axis. We should chose temperature as the explantory variable.

Tessa calculated the linear regression equation as $w = 10\,755 - 171t$.

(e) Give an interpretation of the gradient of this regression equation.

(1)

The gradient is the coefficient of t, in this case it is −171.
The means that for each 1°C temperature increase, sales will fall by £171.

(Total for Question 1 is 7 marks)

2. An ornithologist believes that there is a relationship between the tail length, *t* mm, and the wing length, *w* mm, of female hook-billed kites. A random sample of size 10 is taken from a database of these kites and the relevant data is given in the table below.

***t* (mm)**	191	197	208	180	188	210	196	191	179	208
***w* (mm)**	284	285	288	273	280	283	288	271	257	289

The ornithologist plans to use a linear regression model based on these data and interpolate or extrapolate as necessary to estimate the wing length of other female hook-billed kites from their tail length.

(a) (i) Explain what is meant by extrapolation.

Extrapolation is using the trends in data we have to predict other/future outcomes. **(1)**

(ii) Explain the dangers of extrapolation.

This may change in the future and predictions could be wrong. **(1)**

For example predicated football ticket sales generally increases over time, but due to Covid-19, this trend didn't continue in 2020, and extrapolating wouldn't take this into account.

The ornithologist attempts to calculate the product moment correlation coefficient, *r*, and obtains a value of 1.3.

(b) Explain how she should be able to identify that this is incorrect without carrying out any further calculations.

This is incorrect since $-1 \leq r \leq 1$ since ±1 shows a perfect correlation. **(1)**

(c) Use your calculator to find the correct value of the product moment correlation coefficient, *r*.

From the formula sheet we have that $r = S_{xy} / \sqrt{S_{xx} S_{yy}}$ **(1)**

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$\sum x_i = \sum t_i = 1948$, $\sum y_i = \sum w_i = 2798$

$\sum x_i^2 = \sum t_i^2 = 380600$, $\sum y_i^2 = \sum w_i^2 = 783798$

$\sum x_i y_i = \sum t_i w_i = 545827$

$$S_{xx} = 380600 - \frac{1948^2}{10} = 1129.6$$

$$S_{yy} = 783798 - \frac{2798^2}{10} = 917.6$$

$$S_{xy} = 545827 - \frac{1948 \times 2798}{10} = 776.6$$

$$\Rightarrow \quad r = \frac{776.6}{\sqrt{1129.6 \times 917.6}} = \underline{\underline{0.76}}$$

(d) Stating your hypotheses clearly test, at the 1% significance level, whether or not there is evidence that the product moment correlation coefficient for the population is positive. **(3)**

$H_0 : r = 0$ v.s $H_1 : r > 0$.

We have that $n = 10$, $\alpha = 0.01$ which means that our coefficient is 0.7155 (reading off the table). $0.76 > 0.7155$, which means that r is significant and we reject H_0 and conclude that $r > 0$ and there is a significant positive correlation between the two variables.

(e) Explain what your test in part (d) suggests about female hook-billed kites.

(1)

As their tail length increases, their wing length will increase.

(Total for Question 2 is 8 marks)

3. Sam is investigating the weather throughout the year in Hurn.

He uses the large data set to investigate the daily mean wind direction.

Sam believes that each cardinal wind direction is equally likely in Hurn.

(a) Assuming that Sam is correct,

(i) state the probability that the cardinal wind direction in Hurn on a randomly selected day is NNE,

(1)

There are 16 cardinal wind directions, so we have that the Probability of NNE is $\frac{1}{16} = 0.0625$.

(ii) state the distribution that Sam should use to model the probability of each cardinal wind direction in Hurn on a randomly selected day.

(1)

We have 16 equally likely outcomes, so this data is from the discrete uniform distribution, as it requires n equally likely outcomes.

Sam decides to investigate the daily mean wind direction throughout the year.

(b) State a limitation of using the data for Hurn from the large data set as a sampling frame.

(1)

There may be missing data in the large data set, or it may have additional data from other locations which would skew the results.

(c) Explain how to use simple random sampling to select 36 days from a year.

(2)

let each day in the Year be represented, each, by a number (from 1 to 365), then generate 36 random numbers and pick the corresponding days to be Sampled.

Sam defines the random variable *X* as the number of days out of 36 on which the daily mean wind direction in Hurn is between the bearings 135° and 225°.

Sam collects data from 36 randomly selected days and finds that $x = 15$.

Sam carries out a hypothesis test at the 10% level of significance.

(d) Given that $H_0 : p = 0.25$ and that the critical region is $\{X \leq 4 \cup X \geq 14\}$,

(i) state the alternative hypothesis, H_1

(ii) giving a reason for your answer, explain what Sam should conclude about the daily mean wind direction in Hurn.

(3)

(i) $H_1 : p \neq 0.25$ (note we will have a two-tailed/sided test).

(ii) We can use normal approximation.

$\mu = np = 36 \times 0.25 = 9$ and $\sigma^2 = npq = 36 \times 0.25 \times 0.75 = 6.75 \Rightarrow X \sim N(9, 6.75^2)$

Then $P(X \geqslant 15) = \mathbb{P}\left(Z \geqslant \frac{15-9}{\sqrt{6.75}}\right) = \mathbb{P}(Z \geqslant 2.3) = 1 - \Phi(2.3) = 1 - 0.98928 = 0.011.$

We have a 2 tailed test $\Rightarrow$ P-value $= 2 \times 0.11 = 0.22$.

At $\alpha = 10\% = 0.1$, we have $0.22 > 0.1 \Rightarrow$ We do not reject $H_0 \Rightarrow p = 0.25$.

This means that Sam can conclude that there is a 25% chance of the wind being between 135° and 225°.

(Total for Question 3 is 8 marks)

4. A meteorologist believes that there is a relationship between the daily mean windspeed, *w* kn, and the daily mean temperature, *t* °C. A random sample of 9 consecutive days is taken from past records from a town in the UK in July and the relevant data is given in the table below.

t	13.3	16.2	15.7	16.6	16.3	16.4	19.3	17.1	13.2
w	7	11	8	11	13	8	15	10	11

The meteorologist calculated the product moment correlation coefficient for the 9 days and obtained $r = 0.609$.

(*a*) Explain why a linear regression model based on these data is unreliable on a day when the mean temperature is 24 °C.

None of the Sampled values are near to the mean of 24°C, which would mean the data from these 9 days would be unreliable.

(1)

(*b*) State what is measured by the product moment correlation coefficient.

The product moment correlation coefficient measures how linear the relationship between two variables is.

(1)

(*c*) Stating your hypotheses clearly test, at the 5% significance level, whether or not the product moment correlation coefficient for the population is greater than zero.

$H_0: r=0$ v.s. $H_1: r>0$, then our critical value for a one-sided test at $\alpha = 0.05$ with $n=9$ is 0.5822. We have $r = 0.609 > 0.5822$, which means that r is Significant and we reject H_0 and Conclude that $r>0$.

(3)

Using the same 9 days, a location from the large data set gave $\bar{t} = 27.2$ and $\bar{w} = 3.5$.

(*d*) Using your knowledge of the large data set, suggest, giving your reason, the location that gave rise to these statistics.

This mean temperature is much higher than one would suspect from the UK so we conclude that the data is from abroad; it is from Beijing.

(1)

(Total for Question 4 is 6 marks)

---

5. Barbara is investigating the relationship between average income (GDP per capita), *x* US dollars, and average annual carbon dioxide (CO_2) emissions, *y* tonnes, for different countries.

She takes a random sample of 24 countries and finds the product moment correlation coefficient between average annual CO_2, emissions and average income to be 0.446

(*a*) Stating your hypotheses clearly, test, at the 5% level of significance, whether or not the product moment correlation coefficient for all countries is greater than zero.

(3)

let $H_0: r=0$ v.S. $H_1: r>0$. Then for $n=24, \alpha=0.05$ and a one-sided test we have that our critical value is 0.3438 from the coefficient table. Then $r = 0.446 > 0.3438 \Rightarrow$ we reject H_0 and conclude that $r>0$.

Barbara believes that a non-linear model would be a better fit to the data.
She codes the data using the coding $m = \log_{10} x$ and $c = \log_{10} y$ and obtains the model $c = -1.82 + 0.89m$

The product moment correlation coefficient between c and m is found to be 0.882

(*b*) Explain how this value supports Barbara's belief.

The value of $r = 1$ would mean perfect positive correlation, and the closer the value is to 1, the better the correlation will be.

Linear model gives $0.446 < 0.882$ from the non-linear model. 0.882 is closer to 1 than 0.446 => non-linear model is better fitting for the data.

(1)

(*c*) Show that the relationship between y and x can be written in the form $y = ax^n$ where a and n are constants to be found.

(5)

$\rightarrow y = \log_b(x) \Leftrightarrow b^y = x$

$C = -1.82 + 0.89m \Rightarrow \log_{10}(y) = -1.82 - 0.89\log_{10}(x)$

$\Rightarrow y = 10^{-1.82 - 0.89\log_{10}(x)} \longrightarrow a\log(x) = \log(x)^a$

$y = 10^{-1.82} \cdot 10^{\log(x)^{0.89}}$

$y = 0.015 \cdot 10^{\log(x)^{0.89}} \Leftrightarrow y = \underline{0.015x^{0.89}} \Rightarrow$ $a = 0.015$ and $n = 0.89$.

(Total for Question 5 is 9 marks)

---

6. A company sells seeds and claims that 55% of its pea seeds germinate.

(*a*) Write down a reason why the company should not justify their claim by testing all the pea seeds they produce.

If a population (in this case the number of pea's) is large then it can be challenging and time consuming/expensive to collect and analyse such a large amount of information.

(1)

A random selection of the pea seeds is planted in 10 trays with 24 seeds in each tray.

(*b*) Assuming that the company's claim is correct, calculate the probability that in at least half of the trays 15 or more of the seeds germinate.

For one tray we have that $X \sim \text{Binom}(24, 0.55)$ where X is a random variable which tells us the number of seeds that germinate in each tray.

We want the probability that 15 or more seeds germinate, i.e. $P(X \geqslant 15)$.

where $x = 15$, $n = 24$ and $p = 0.55$. $P(X \geqslant 15) = 0.2991$.

Let us then repeat this and use another random variable Y, to work out the number of trays where 15 or more seeds germinate. We have $n = 10$ (10 trays) and $p = 0.2992$. We want to check if at least half of the trays see the seeds germinate.

(3)

$\Rightarrow Y \sim B(10, 0.2992)$

$\Rightarrow \mathbb{P}(X \geqslant 5) = \underline{0.1487}$

(*c*) Write down two conditions under which the normal distribution may be used as an approximation to the binomial distribution.

We can use normal approximation to the binomial distribution if

- n is large and • p is close to or equal to $1/2$.

(1)

A random sample of 240 pea seeds was planted and 150 of these seeds germinated

(*d*) Assuming that the company's claim is correct, use a normal approximation to find the probability that at least 150 pea seeds germinate.

(3)

Normal approximation : $p = 0.55$, $q = 1-p = 0.45$ and $n = 240$, $x = 150$

$\Rightarrow X \sim N(np, npq) = N(\overset{\mu}{132}, \overset{\sigma^2}{59.4}) \Rightarrow P(x \geqslant 15)$ $\quad P\left(z \geqslant \frac{150-132}{\sqrt{59.4}}\right) = P(z > 2.34)$

$\Rightarrow 1 - \Phi(2.34) = 1 - 0.9910 = 0.009$ from Normal distribution table.

$\Rightarrow$ our probability of $\geqslant 150$ seeds germinating is $\underline{\underline{0.009}}$.

(*e*) Using your answer to part (*d*), comment on whether or not the proportion of the company's pea seeds that germinate is different from the company's claim of 55%

(1)

let $H_0 : p = 0.55$ v.s. $H_1 : p \neq 0.55$.

our p-value is $2 \times 0.009 = 0.018$, noting we multiply by two as we are now carrying out a two sided test to see if the proportion of seeds is <u>different</u> (as opposed to greater).

Then for $\alpha = 0.05$ or 5% significance level, $0.05 > 0.018$ which means that we reject H_0 in favour of H_1 and conclude that $p \neq 0.55$.

Note that we would also reject H_0 and accept $H_1 : p > 0.55$ for a one-tailed test with p-value $\underline{\underline{0.009}}$.

(Total for Question 6 is 9 marks)

7. The lifetime, L hours, of a battery has a normal distribution with mean 18 hours and standard deviation 4 hours.

Alice's calculator requires 4 batteries and will stop working when any one battery reaches the end of its lifetime.

(a) Find the probability that a randomly selected battery will last for longer than 16 hours. **(1)**

We want to find the probability that a randomly Selected battery will last longer than 16 hours $\iff P(L>16)$.

We have that $\mu = 18$ and $\sigma = 4$ hours. $\Rightarrow P(L>16) = P\left(Z > \frac{16-18}{4}\right) = P(Z>-0.5)$

Then we have that $P(Z>-0.5) = \Phi(0.5) = \underline{0.6915}$.

At the start of her exams Alice put 4 new batteries in her calculator. She has used her calculator for 16 hours, but has another 4 hours of exams to sit.

(b) Find the probability that her calculator will not stop working for Alice's remaining exams. **(5)**

We want to work out the probability that th calculator will work for 20 hours given that it has lasted for 6, i.e. $P(L>20 \mid L>16)$.

We can use Bayes Theorem : $P(L>20 \mid L>16) = \frac{P(L>16 \mid L>20) \times P(L>20)}{1(L>16)}$

We have that $P(L>16)$ is 0.6915 from (a).

$P(L>20)$ can be found again using area under normal distribution curve $\Rightarrow P(L>20) = 1 \; \Phi(0.5) = \underline{0.3085}$.

Then $P(L>16 \mid L>20)$ is 1, we know if $L>20$ then $L>16$ since $20>16$.

$\Rightarrow P(L>20 \mid L>16) = \frac{1 \times 0.3085}{0.6915} = 0.446$ per battery

Then the Probability that the calculator will stop working is $(0.446)^4 = \underline{0.04}$

Alice only has 2 new batteries so, after the first 16 hours of her exams, although her calculator is still working, she randomly selects 2 of the batteries from her calculator and replaces these with the 2 new batteries.

(c) Show that the probability that her calculator will not stop working for the remainder of her exams is 0.199 to 3 significant figures. **(3)**

We want to multiply $P(L>20)^2$ (as we have 2 batteries under that probability, and then we must find $P(L>4)^2$ as we replace 2 batteries with 4 hours to go.

$P(L>4) = 0.9998$.

Then the probability that her calculator will n t op working is

$(0.9998)^2 (0.446)^2 = \underline{0.198}$

After her exams, Alice believed that the lifetime of the batteries was more than 18 hours. She took a random sample of 20 of these batteries and found that their mean lifetime was 19.2 hours.

(d) Stating your hypotheses clearly and using a 5% level of significance, test Alice's belief.

(5)

$H_0 : \mu = 18$ v.s. $H_1 : \mu > 18$

Then $P(L > 19.2) = P\left(Z > \frac{19.2 - 18}{4}\right) = P(Z > 0.3) = 1 - \Phi(0.3) = 0.61$

(Total for Question 7 is 14 marks)