# Statistics and Probability\_P\_3

**1a.** *[1 mark]*

Two IB schools, A and B, follow the IB Diploma Programme but have different teaching methods. A research group tested whether the different teaching methods lead to a similar final result.

For the test, a group of eight students were randomly selected from each school. Both samples were given a standardized test at the start of the course and a prediction for total IB points was made based on that test; this was then compared to their points total at the end of the course.

Previous results indicate that both the predictions from the standardized tests and the final IB points can be modelled by a normal distribution.

It can be assumed that:

* the standardized test is a valid method for predicting the final IB points
* that variations from the prediction can be explained through the circumstances of the student or school.

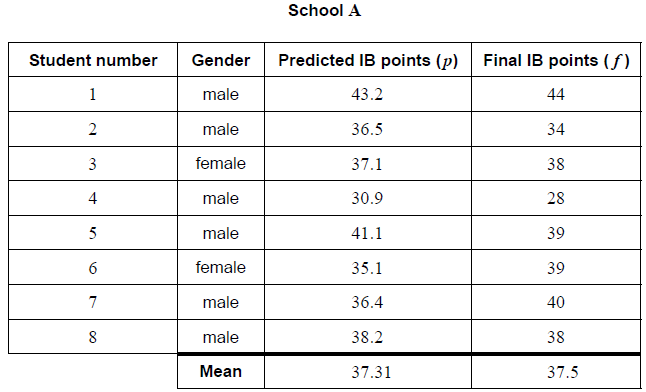
Identify a test that might have been used to verify the null hypothesis that the predictions from the standardized test can be modelled by a normal distribution.

**1b.** *[1 mark]*

State why comparing only the final IB points of the students from the two schools would not be a valid test for the effectiveness of the two different teaching methods.

**1c.** *[1 mark]*

The data for school A is shown in the following table.



For each student, the change from the predicted points to the final points  was calculated.

Find the mean change.

**1d.** *[2 marks]*

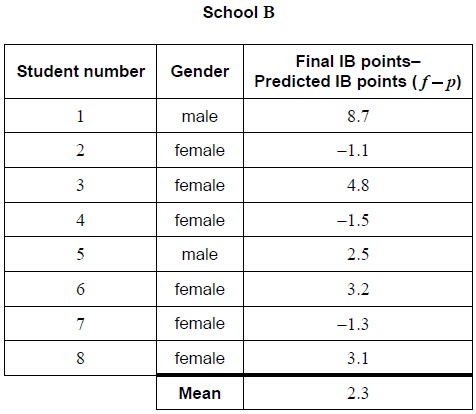
Find the standard deviation of the changes.

**1e.** *[4 marks]*

Use a paired -test to determine whether there is significant evidence that the students in school A have improved their IB points since the start of the course.

**1f.** *[5 marks]*

The data for school B is shown in the following table.



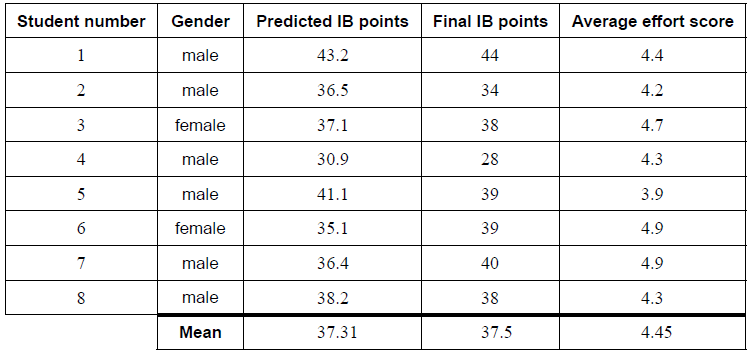
Use an appropriate test to determine whether there is evidence, at the 5 % significance level, that the students in school B have improved more than those in school A.

**1g.** *[1 mark]*

State why it was important to test that both sets of points were normally distributed.

**1h.** *[3 marks]*

School A also gives each student a score for effort in each subject. This effort score is based on a scale of 1 to 5 where 5 is regarded as outstanding effort.



It is claimed that the effort put in by a student is an important factor in improving upon their predicted IB points.

Perform a test on the data from school A to show it is reasonable to assume a linear relationship between effort scores and improvements in IB points. You may assume effort scores follow a normal distribution.

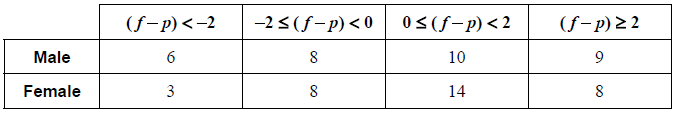
**1i.** *[1 mark]*

Hence, find the expected improvement between predicted and final points for an increase of one unit in effort grades, giving your answer to one decimal place.

**1j.** *[6 marks]*

A mathematics teacher in school A claims that the comparison between the two schools is not valid because the sample for school B contained mainly girls and that for school A, mainly boys. She believes that girls are likely to show a greater improvement from their predicted points to their final points.

She collects more data from other schools, asking them to class their results into four categories as shown in the following table.



Use an appropriate test to determine whether showing an improvement is independent of gender.

**1k.** *[2 marks]*

If you were to repeat the test performed in part (e) intending to compare the quality of the teaching between the two schools, suggest **two** ways in which you might choose your sample to improve the validity of the test.

**2a.** *[3 marks]*

This question will connect Markov chains and directed graphs.

Abi is playing a game that involves a fair coin with heads on one side and tails on the other, together with two tokens, one with a fish’s head on it and one with a fish’s tail on it. She starts off with no tokens and wishes to win them both. On each turn she tosses the coin, if she gets a head she can claim the fish’s head token, provided that she does not have it already and if she gets a tail she can claim the fish’s tail token, provided she does not have it already. There are 4 states to describe the tokens in her possession; A: no tokens, B: only a fish’s head token, C: only a fish’s tail token, D: both tokens. So for example if she is in state B and tosses a tail she moves to state D, whereas if she tosses a head she remains in state B.

Draw a transition state diagram for this Markov chain problem.

**2b.** *[1 mark]*

Explain why for any transition state diagram the sum of the out degrees of the directed edges from a vertex (state) must add up to +1.

**2c.** *[3 marks]*

Write down the transition matrix **M**, for this Markov chain problem.

**2d.** *[4 marks]*

Find the steady state probability vector for this Markov chain problem.

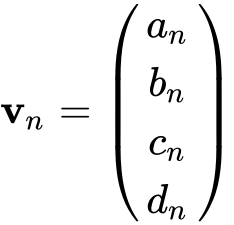
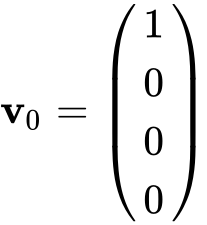
**2e.** *[1 mark]*

Explain which part of the transition state diagram confirms this.

**2f.** *[2 marks]*

Explain why having a steady state probability vector means that the matrix **M** must have an eigenvalue of .

**2g.** *[4 marks]*

After  throws the probability vector, for the 4 states, is given by  where the numbers represent the probability of being in that particular state, e.g.  is the probability of being in state B after  throws. Initially .

Find .

**2h.** *[2 marks]*

Hence, deduce the form of .

**2i.** *[2 marks]*

Explain how your answer to part (f) fits with your answer to part (c).

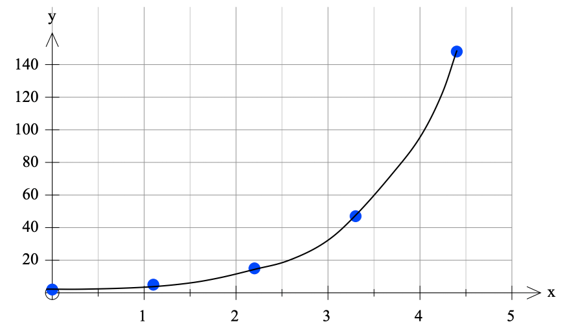
**2j.** *[4 marks]*

Find the minimum number of tosses of the coin that Abi will have to make to be at least 95% certain of having finished the game by reaching state C.

**3a.** *[3 marks]*

*This question explores methods to determine the area bounded by an unknown curve.*

The curve  is shown in the graph, for .



The curve  passes through the following points.



It is required to find the area bounded by the curve, the -axis, the -axis and the line .

Use the trapezoidal rule to find an estimate for the area.

**3b.** *[2 marks]*

With reference to the shape of the graph, explain whether your answer to part (a)(i) will be an over-estimate or an underestimate of the area.

**3c.** *[3 marks]*

One possible model for the curve  is a cubic function.

Use all the coordinates in the table to find the equation of the least squares cubic regression curve.

**3d.** *[1 mark]*

Write down the coefficient of determination.

**3e.** *[2 marks]*

Write down an expression for the area enclosed by the cubic function, the -axis, the -axis and the line .

**3f.** *[2 marks]*

Find the value of this area.

**3g.** *[2 marks]*

A second possible model for the curve  is an exponential function, , where .

Show that .

**3h.** *[1 mark]*

Hence explain how a straight line graph could be drawn using the coordinates in the table.

**3i.** *[5 marks]*

By finding the equation of a suitable regression line, show that  and .

**3j.** *[2 marks]*

Hence find the area enclosed by the exponential function, the -axis, the -axis and the line .

**4a.** *[1 mark]*

*This question explores methods to analyse the scores in an exam.*

A random sample of 149 scores for a university exam are given in the table.



Find unbiased estimates for the population mean.

**4b.** *[2 marks]*

Find unbiased estimates for the population Variance.

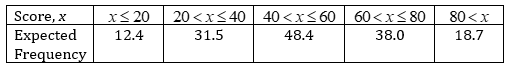
**4c.** *[3 marks]*

The university wants to know if the scores follow a normal distribution, with the mean and variance found in part (a).

Show that the expected frequency for 20 <  ≤ 4 is 31.5 correct to 1 decimal place.

**4d.** *[8 marks]*

The expected frequencies are given in the table.



Perform a suitable test, at the 5% significance level, to determine if the scores follow a normal distribution, with the mean and variance found in part (a). You should clearly state your hypotheses, the degrees of freedom, the *p*-value and your conclusion.

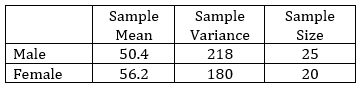
**4e.** *[2 marks]*

The university assigns a pass grade to students whose scores are in the top 80%.

Use the normal distribution model to find the score required to pass.

**4f.** *[6 marks]*

The university also wants to know if the exam is gender neutral. They obtain random samples of scores for male and female students. The mean, sample variance and sample size are shown in the table.



Perform a suitable test, at the 5% significance level, to determine if there is a difference between the mean scores of males and females. You should clearly state your hypotheses, the *p*-value and your conclusion.

**4g.** *[6 marks]*

The university awards a distinction to students who achieve high scores in the exam. Typically, 15% of students achieve a distinction. A new exam is trialed with a random selection of students on the course. 5 out of 20 students achieve a distinction.

Perform a suitable test, at the 5% significance level, to determine if it is easier to achieve a distinction on the new exam. You should clearly state your hypotheses, the critical region and your conclusion.

**4h.** *[3 marks]*

A different exam is trialed with 16 students. Let  be the percentage of students achieving a distinction. It is desired to test the hypotheses

 against 

It is decided to reject the null hypothesis if the number of students achieving a distinction is greater than 3.

Find the probability of making a Type I error.

**4i.** *[3 marks]*

Given that  find the probability of making a Type II error.

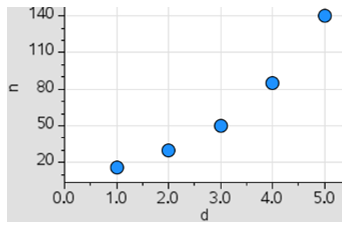
**5a.** *[3 marks]*

*In this question you will explore possible models for the spread of an infectious disease*

An infectious disease has begun spreading in a country. The National Disease Control Centre (NDCC) has compiled the following data after receiving alerts from hospitals.



A graph of  against  is shown below.



The NDCC want to find a model to predict the total number of people infected, so they can plan for medicine and hospital facilities. After looking at the data, they think an exponential function in the form  could be used as a model.

Use an exponential regression to find the value of  and of , correct to 4 decimal places.

**5b.** *[3 marks]*

Use your answer to part (a) to predict

the number of new people infected on day 6.

**5c.** *[2 marks]*

the day when the total number of people infected will be greater than 1000.

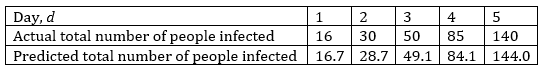
**5d.** *[1 mark]*

The NDCC want to verify the accuracy of these predictions. They decide to perform a  goodness of fit test.

Use your answer to part (a) to show that the model predicts 16.7 people will be infected on the first day.

**5e.** *[2 marks]*

The predictions given by the model for the first five days are shown in the table.



Explain why the number of degrees of freedom is 2.

**5f.** *[5 marks]*

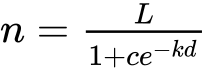
Perform a  goodness of fit test at the 5% significance level. You should clearly state your hypotheses, the p-value, and your conclusion.

**5g.** *[2 marks]*

In fact, the first day when the total number of people infected is greater than 1000 is day 14, when a total of 1015 people are infected.

Give two reasons why the prediction in part (b)(ii) might be lower than 14.

**5h.** *[2 marks]*

Based on this new data, the NDCC decide to try a logistic model in the form .

Use the data from days 1–5, together with day 14, to find the value of

.

**5i.** *[1 mark]*

.

**5j.** *[1 mark]*

.

**5k.** *[2 marks]*

Hence predict the total number of people infected by this disease after several months.

**5l.** *[3 marks]*

Use the logistic model to find the day when the rate of increase of people infected is greatest.