Section 2: Testing for correlation

Notes and Examples

These notes contain subsections on

Correlation coefficients

A correlation coefficient is a measure of correlation. It takes values between -1 and 1, with -1 representing perfect negative correlation, 0 representing no correlation, and 1 representing perfect positive correlation.

There are different types of correlation coefficient for different situations. A correlation coefficient may measure linear correlation, which is an indication of how close the data lie to a straight line. In this case, a correlation of 1 means that all the data lie on a straight line with positive gradient, and a correlation of -1 means that all the data lie on a straight line with negative gradient.

Other correlation coefficients measure other types of association, rather than linear correlation. For example, sometimes you might be interested in rank correlation, which compares the rank order of two sets of data. An example might be if two judges rank cakes baked by 10 different people. In that case, a correlation of 1 means that both judges have placed the cakes in exactly the same order.

Hypothesis test involving correlation coefficients

To carry out a hypothesis test for correlation or association, you must first state the null and alternative hypothesis.

The null hypothesis has the form

H_0: $\rho = 0$ where ρ is the correlation coefficient for the parent population

The alternative hypothesis takes one of the forms

H_1: $\rho > 0$ (for a one-tailed test testing for positive correlation)
H_1: $\rho < 0$ (for a one-tailed test testing for negative correlation)
H_1: $\rho \neq 0$ (for a two-tailed test)

The correlation coefficient for the sample data is compared with a critical value, which depends on the number of data items in the sample, the significance level, and whether the test is one-tailed or two-tailed. You look up the critical value in tables.

If the correlation coefficient is higher than the critical value (in the case of negative correlation coefficients, compare the absolute value of the correlation coefficient) then the null hypothesis is rejected. There is evidence to suggest that there is correlation.

Otherwise, the null hypothesis is accepted. There is not sufficient evidence to suggest that there is correlation.

Example 1
A researcher wishes to find out if there is any connection between the length of time young children spend using computers and their reading ability.
She collects data on 50 seven-year-olds, which gives a correlation coefficient of -0.3608
Carry out a hypothesis test at the 5% level to determine whether there is any connection between time spent on computers and reading ability.

Solution
$H_0 : \rho = 0$
$H_1 : \rho \neq 0$
where ρ is the correlation coefficient for the parent population.

From tables, the critical value for a two-tailed test with $n = 50$ at the 5% significance level is 0.2787.
Since 0.3608 > 0.2787, reject H_0. There is evidence to suggest that there is correlation between time spent using computers and reading ability.

Note the following important points about the example above.

- The test is two-tailed (H_1 is $\rho \neq 0$) because of the wording of the problem. The researcher might want to look at whether using a computer improves children's reading ability (in which case the test would be one-tailed, with alternative hypothesis $\rho > 0$), or she might have wanted to investigate whether using a computer has a detrimental effect on children's reading (again a one-tailed test, but this time with alternative hypothesis $\rho < 0$). As it is, she was investigating whether there is any correlation at all.

- The correlation coefficient is negative. You need to disregard the sign of the correlation coefficient when comparing with the critical value.

- The conclusion is that there is evidence to suggest a **correlation** between the two variables. It does **not** mean that using computers has a negative effect on a child's reading ability. For example, it might be the case that children who read well spend less time on a computer because they enjoy reading books, or perhaps parents who spend time helping their children with reading may also be those who are more likely to restrict the time their children spend on the computer. Issues such as these are very complex, and all that the hypothesis test tells us is that there appears to be some connection between the two variables.

Edexcel A level Hypothesis tests 2 Notes & Examples

Using a p-value

Sometimes you may be given a p-value instead of a value for the correlation coefficient. You have used p-values in hypothesis tests previously. In this context, the p-value is the probability that, if there is no correlation, a random sample gives the given value for the correlation coefficient – in other words, the probability that this correlation coefficient could have been obtained by chance. Just as in other types of hypothesis test, you compare the p-value with the significance level. If the p-value is less than the significance level, this means that it is very unlikely that this value for the correlation coefficient could have arisen by chance, so the null hypothesis is rejected.

