Statistics Sector 1: Hypothesis Tests - Correlation

Definitions

Null Hypothesis (H_0):
an assertion that a parameter in a statistical model takes a particular value, and is assumed true until experimental evidence suggests otherwise.

Alternative Hypothesis (H_1):
expresses the way in which the value of a parameter may deviate from that specified in the null hypothesis, and is assumed true when the experimental evidence suggests that the null hypothesis is false.

Test statistic:
a function of a sample of observations which provides a basis for testing the validity of the null hypothesis.

Critical region:
the null hypothesis is rejected when a calculated value of the test statistic lies within this region.

Critical value:
the value which determines the boundary of the critical region.

Significance level:
the size of the critical region

One-tailed test:
the critical region is located wholly at one end of the sampling distribution of the test statistic.
H_1 involves < or > but not both.

Two-tailed test:
the critical region comprises areas at both ends of the sampling distribution of the test statistic.
H_1 involves $\neq$.

Testing a Hypothesis about Product Moment Correlation Coefficient with Critical Values

1) State the hypothesis

   $H_o: \rho = 0$ (no correlation exists in the population between the two variable)
   $H_1: \rho < 0, \rho > 0 \; or \; \rho \neq 0$ (population correlation coefficient is less than, more than or not equal to zero)

2) Test Statistic

   This is the value of r.

3) Critical Value

   This will be given in the question.

4) Conclusion

   Conclude in the context of the question, if $|r| > |CV|$ reject H_0.

Example 1

A technician monitoring water purity believes that there is a relationship between the hardness of the water and its alkalinity. Over a period of 10 days, the technician recorded the alkalinity and hardness (mg/l) and calculated $r = 0.9264$ to 4 decimal places. The critical value is ± 0.5494.

a) Test, at the 5% level the hypothesis that higher alkalinity is associated with higher water hardness, stating your conclusions clearly.

Hypothesis

$H_0: \rho = 0$

$H_1: \rho > 0$

Test Statistic

$r = 0.9264$

Critical Value

$CV = \pm 0.5494$

Conclusion

$0.9264 > 0.5494$ ∴ Reject H_0

There is significant evidence to suggest that there is positive correlation between higher alkalinity and higher water hardness.

b) What assumption about the sample have you had to make in order to be able to carry out this hypothesis test?

The sample must be random.

Example 2

The data for 20 students gave a correlation coefficient of $r = -0.8$, between time spent socialising and test results. The critical value is ± 0.3783.Does this support a claim that there is negative correlation at the 5% significance level?

Hypothesis

$H_0: \rho = 0$

$H_1: \rho < 0$

Test Statistic

$r = -0.8$

Critical Value

$C.V = \pm 0.3783$

Conclusion

$0.8 > 0.3783$ ∴ Reject H_0

There is significant evidence to suggest there is negative correlation between time spent socialising and test results.

Example 3

The results given show the yield, y, in grams from a chemical experiment corresponding to a input of x grams of a chemical.

x	5.6	6.3	8.5	4.2	7.4	5.1	9.6	4.8	6.9	5.9
y	82	78	86	65	91	80	75	72	89	74

Given that $r = 0.484$ and the critical value is ± 0.6319 investigate whether there is any correlation between the input and the yield. Use a 5% significance level.

Hypothesis

$H_0: \rho = 0$

$H_1: \rho \neq 0$

Test Statistic

$r = 0.484$

Critical Value

$C.V = \pm 0.6319$

Conclusion

$0.484 < 0.6319$ ∴ Accept H_0

There is no significant evidence to suggest there is any correlation between the input and the yield.

Testing a Hypothesis about Product Moment Correlation Coefficient with p-values

So far in making decisions in hypothesis tests critical values have been used. There is another approach using p values. Often stats packages give you p-values and in many research papers p-values are quoted.

1) State the hypothesis

   $H_o: \rho = 0$ (no correlation exists in the population between the two variable)
   $H_1: \rho < 0, \rho > 0 \; or \; \rho \neq 0$ (population correlation coefficient is less than, more than or not equal to zero)

2) Conclusion

   Conclude in the context of the question, if the p-value is less than the significance level reject H_0.

Example 4

Data on the number of hours spent training for a triathlon and the number of hours taken to complete the triathlon. It is suspect that there is a negative correlation between training time and triathlon time. The data was inputted into a stats package and the following output was obtained, investigate the claim at the 10% significance level.

Correlation: x, y

```
Pearson correlation of x and y = -0.287
P-Value = 0.422
```

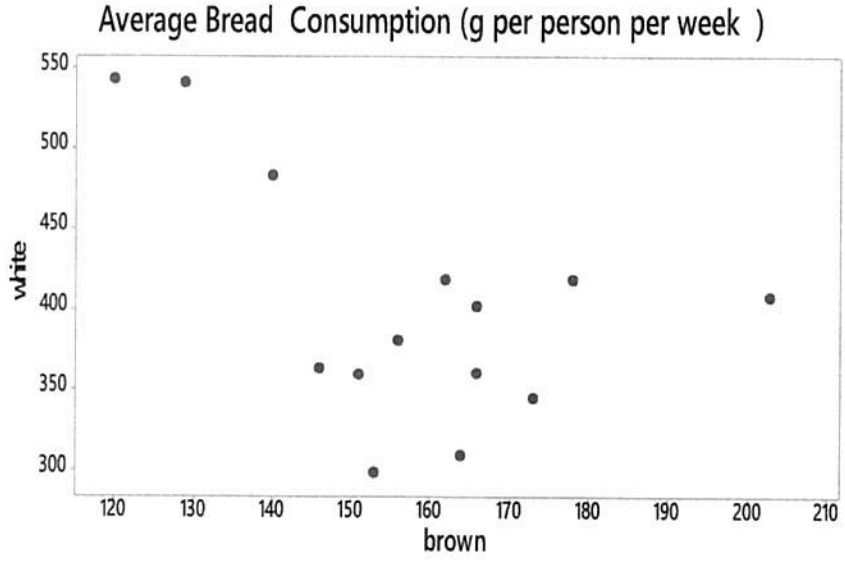
$H_0: \rho = 0$

$H_1: \rho < 0$

$0.422 > 0.10 \therefore$ Accept H_0

There is no significant evidence to suggest that there is a negative ~~amount~~ correlation between the numbe of hours spent training and the number of hours taken to complete a triathlon.

Example 5



Correlation: white, brown

```
Pearson correlation of white and brown = -0.517
P-Value = 0.058
```

Investigate the claim that there is some correlation between the average consumption of white and brown bread per person per week, at the 2% significance level.

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$0.058 > 0.02 \therefore$ Accept H_0

There is no significant evidence to suggest there is any correlation between average consumption of brown and white bread.

Exam Style Questions

1) The average daily temperature in a town and the number of newspapers sold by shops in that town are measured over a year to see if there is any correlation.

a State the hypotheses being tested. [2 marks]

$H_0: \rho = 0$

$H_1: \rho \neq 0$

b Given that the pmcc is 0.0802 and the critical value is ± 0.0851, determine the conclusion to the test in context. [3 marks]

$r = 0.0802 \quad C.V = \pm 0.0851$

$0.0802 < 0.0851 \therefore$ Accept H_0

There is no significant evidence to suggest that there is any correlation between average daily temperature and numbe of newspapers sold by shops

c The sample size is greatly increased but the same pmcc is obtained. What would you expect the conclusion of the new test to be? [2 marks]

As the sample size increases the critical value will decrease so the conclusion for the new test will be to reject H_0, there is significant evidence to suggest correlation between average daily temp and numbe of newspapes sold.

2) a The scores obtained in a maths test by a class of students, along with the number of hours they revised for the test, are measured for correlation. A test is performed at the 5% significance level and the pmcc has a *p*-value of 4.74%

Determine the conclusion to the test in context. [3 marks]

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$0.0474 < 0.05 \therefore$ Reject H_0

There is significant evidence to suggest a correlation between scores obtained in a maths class and numbe of hours they revised.

b A student decides that, based on the result of this hypothesis test, they should revise for more hours. Why might this logical step not be supported by the result of the hypothesis test? [1 mark]

Correlation does not necessarily imply causation.

3) The number of ice cream tubs eaten per person per week in the North East and the North West are tested for positive correlation. A sample of 14 years is taken and the data have a pmcc of 0.519. The critical value is 0.458 at the 5% significance level. Perform a hypothesis test for correlation, stating your hypothesis clearly. [5 marks]

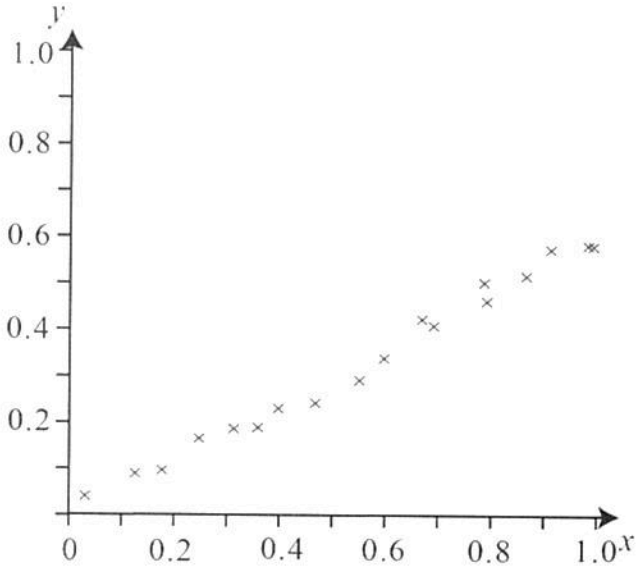
$H_0: \rho = 0$
$H_1: \rho > 0$

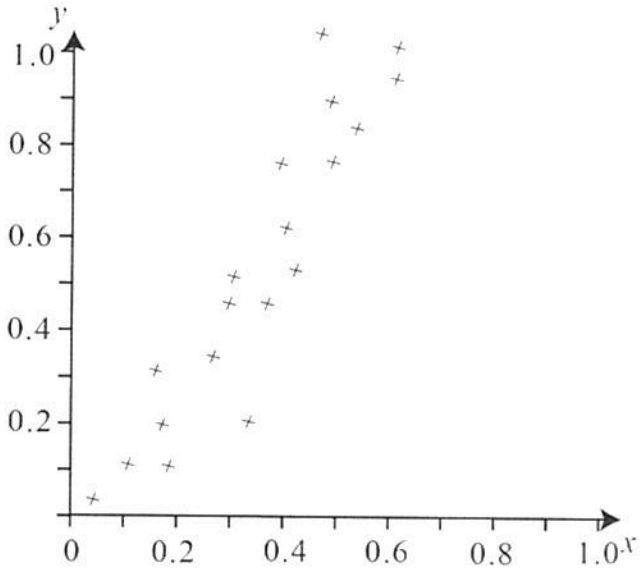
Test Statistic $r = 0.519$
Critical Value $C.V = 0.458$

$0.519 > 0.458$ ∴ Reject H_0

There is significant evidence to suggest the number of ice cream tubs eaten per person per week in the North East and North West has positive correlation.

4) These two scatter diagrams show data for two different situations.





a Someone claims that the diagram on the left displays data with a stronger correlation because the gradient is steeper. Explain why they are wrong. [2 marks]

The gradient will only tell you if correlation is postive or negative, not the streagth. Correlation tells you how close to a straight line the points are.

b With reference to the pmcc, explain why the diagram on the right is less likely to provide evidence to reject the null hypothesis in a test for correlation. [2 marks]

The diagram on the right has a smaller pmcc (weaker correlation) so it is less likely to be bigge than the critical value which is ~~reje~~ required to reject H_0.