2.3 Further Statistics: Chi-Squared Tests (Goodness of Fit and Contingency Tables): 1-hour Pack

Calculator allowed · about 60 minutes

Total Marks

Name: _____ Date: ____ / ____ / _____

Answer ALL questions. Show all your working.

Key facts: chi-squared tests

Original content written for Revision Library.

A chi-squared test compares observed frequencies O with the expected frequencies E predicted by a proposed model, using the test statistic X^2 = sum of (O - E)^2 / E, calculated over all classes (or cells) after any pooling. For a goodness-of-fit test to a fully specified distribution (for example a uniform distribution, or given proportions), degrees of freedom = (number of classes) - 1. If one or more parameters of the model (for example p for a binomial, or the mean for a Poisson) are estimated from the same sample used in the test, one further degree of freedom is lost for each parameter estimated. For an r by c contingency table testing independence between two variables, expected frequencies are E = (row total * column total) / grand total, and degrees of freedom = (r - 1)(c - 1). Classes with an expected frequency below 5 should be pooled (combined with a neighbouring class) before the test statistic is calculated, reducing the number of classes, and hence the degrees of freedom, accordingly. For any test with 1 degree of freedom, including every 2 by 2 contingency table, Yates' continuity correction is applied: X^2 = sum of (|O - E| - 0.5)^2 / E. The calculated test statistic is compared with the critical value from chi-squared tables at the chosen significance level and the stated degrees of freedom; if the test statistic exceeds the critical value, the null hypothesis is rejected.

1 A researcher records the results of 157 people classified by group and by outcome, as shown. Rows: Group X: Grade A 44, Grade B 23, Grade C 13; Group Y: Grade A 13, Grade B 26, Grade C 38. A test is carried out at the 5% level to see whether the outcome is independent of the group.

	Grade A	Grade B	Grade C
Group X	44	23	13
Group Y	13	26	38

(a) State the hypotheses, and calculate the expected frequency for the first cell in each row. **(3)**

(b) Find the value of the test statistic, to 3 significant figures, and the number of degrees of freedom. **(3)**

(c) Given that the critical value is 5.991, state the conclusion of the test in context. **(3)**

(Total for Question 1 is 9 marks)

© Revision Library | revisionlibrary.co.uk

2 A researcher counts the number of events per interval over 85 intervals: 0 events 15 times, 1 event 29 times, 2 events 23 times, and 3 or more events 18 times. It is suggested that the number of events per interval follows a Poisson distribution. Taking the mean of the data to be 1.518 (treating '3 or more' as 3), test the suggestion at the 5% level.

(a) Calculate the expected frequencies for a Poisson distribution with the given mean, using the categories 0, 1, 2 and 3 or more. **(3)**

(b) Calculate the test statistic and the degrees of freedom, and state the conclusion at the 5% level of significance, given that the critical value is 5.991. **(4)**

(Total for Question 2 is 7 marks)

---

3 A cafe in Norwich records the number of coffees sold in each of four two-hour periods during a working day, over a sample of 180 coffees sold in total, to test whether sales are uniformly distributed across the four periods.

Period 1 (8am-10am): 38. Period 2 (10am-12pm): 52. Period 3 (12pm-2pm): 61. Period 4 (2pm-4pm): 29.

Critical values for chi-squared with 3 degrees of freedom: 10% level 6.251; 5% level 7.815; 2.5% level 9.348; 1% level 11.345; 0.1% level 16.266.

(a) State suitable null and alternative hypotheses for this test. **(2)**

(b) State the expected number of coffees sold in each period under H0, and calculate the chi-squared test statistic. **(3)**

(c) Using the critical values given, state the range within which the p-value of this test lies, justifying your answer. **(2)**

(d) Hence state your conclusion at the 1% significance level, in the context of this question. **(2)**

(Total for Question 3 is 9 marks)

© Revision Library | revisionlibrary.co.uk

4 A GP surgery in Cardiff investigates whether there is an association between having a flu vaccine and catching flu during the winter, among 160 randomly chosen patients. The results are shown in the table below.

	Caught flu	Did not catch flu	Row total
Vaccinated	18	72	90
Not vaccinated	22	48	70
Column totals	40	120	160

(a) State suitable null and alternative hypotheses for this test. **(2)**

(b) Show that the expected frequency for vaccinated patients who caught flu is 22.5, and complete the expected frequency table. **(3)**

(c) Explain why Yates' continuity correction should be used in this test, and calculate the value of chi-squared with the correction applied. **(5)**

(d) State the number of degrees of freedom and the critical value at the 5% significance level, and test whether there is an association between vaccination and catching flu. State your conclusion clearly, in context. **(3)**

(Total for Question 4 is 13 marks)

© Revision Library | revisionlibrary.co.uk

5 A students' union at a university in Sheffield investigates whether attending an exam workshop is associated with passing a resit exam, among 120 students. The results are shown below.

Attended workshop: passed resit 42, failed resit 18, row total 60.
Did not attend: passed resit 24, failed resit 36, row total 60.
Column totals: passed resit 66, failed resit 54, grand total 120.

	Passed resit	Failed resit	Total
Attended workshop	42	18	60
Did not attend	24	36	60
Total	66	54	120

(a) State suitable null and alternative hypotheses for this test. **(2)**

(b) Show that the expected frequency for students who attended the workshop and passed the resit is 33, and complete the expected frequency table. **(3)**

(c) Calculate the chi-squared test statistic, applying Yates' continuity correction. **(3)**

(d) State the number of degrees of freedom, and determine the lowest of the standard significance levels (10%, 5%, 1%) at which the null hypothesis would be rejected. **(3)**

(e) State your conclusion in context, and give one reason why this result does not necessarily mean that attending the workshop causes students to pass the resit. **(2)**

(Total for Question 5 is 13 marks)

© Revision Library | revisionlibrary.co.uk

Mark scheme · 2.3 Further Statistics: Chi-Squared Tests (Goodness of Fit and Contingency Tables): 1-hour Pack

Question 1

(a) H0: outcome and group are independent; H1: they are associated **B1**

(a) expected = (row total x column total)/grand total **M1**

(a) E_{11} = 4560/157 = 29.04, E_{21} = 4389/157 = 27.96 **A1**

(a) Answer: E_{11} = 29.04, E_{21} = 27.96

(b) sum of $(O - E)^2/E$ over all six cells **M1**

(b) chi^2 = 29.3 **A1**

(b) degrees of freedom = (2 - 1)(3 - 1) = 2 **B1**

(b) Answer: chi^2 = 29.3, nu = 2

(c) compares 29.3 with 5.991 **M1**

(c) reject H0 **A1**

(c) there is evidence of an association between group and outcome **B1**

(c) Answer: reject H0

Question 2

(a) $P(X = j) = e^{-1.518}\ 1.518^j/j!$ for j = 0, 1, 2 **M1**

(a) $P(X \geq 3) = 1 - P(X \leq 2)$ **M1**

(a) expected: 18.63, 28.28, 21.46, 16.63 **A1**

(a) Answer: 18.63, 28.28, 21.46, 16.63

(b) chi^2 = sum of $(O - E)^2/E$ **M1**

(b) 0.947 **A1**

(b) degrees of freedom = 4 - 1 - 1 = 2 (the mean was estimated from the data) **B1**

(b) do not reject H0: the Poisson model is a suitable fit **A1**

(b) Answer: chi^2 = 0.947, nu = 2, do not reject H0

Question 3

(a) H0: coffee sales are uniformly distributed across the four periods **B1**

(a) H1: sales are not uniformly distributed across the four periods **B1**

(a) Answer: H0: sales are uniform across the four periods. H1: they are not.

(b) E = 45 for each period **B1**

(b) attempts sum of $(O-E)^2/E$ across all 4 periods **M1**

(b) cao, awrt 13.6 **A1**

(b) Answer: Test statistic = 610/45 = awrt 13.6

(c) compares 13.56 with the 1% (11.345) and 0.1% (16.266) critical values, showing 13.56 exceeds 11.345 but is less than 16.266 **M1**

(c) correct bracket: 0.001 < p < 0.01 **A1**

(c) Answer: 0.001 < p < 0.01

(d) 13.56 > 11.345, so reject H0 at the 1% level **B1**

(d) conclusion in context: there is significant evidence that coffee sales are not uniformly distributed across the four periods **B1**

(d) Answer: Reject H0 at the 1% level: sales are not uniform across the four periods.

Question 4

(a) H0: whether a patient catches flu is independent of whether they were vaccinated **B1**

(a) H1: catching flu is associated with (not independent of) vaccination status **B1**

(a) Answer: H0: catching flu is independent of vaccination status. H1: they are associated.

(b) uses E = (row total * column total) / grand total = 90*40/160 **M1**

(b) 22.5 cso **A1**

(b) remaining three values correct: 67.5, 17.5, 52.5 **A1**

(b) Answer: Expected frequencies: 22.5, 67.5, 17.5, 52.5.

(c) the table is 2 by 2, giving 1 degree of freedom, so Yates' correction is applied to improve the approximation of the discrete data by the continuous chi-squared distribution **B1**

(c) uses X^2 = sum of $(|O-E| - 0.5)^2 / E$ **M1**

(c) first two terms correct (unsimplified): 16/22.5, 16/67.5 **A1**

(c) remaining two terms correct (unsimplified): 16/17.5, 16/52.5 **A1**

(c) cao, awrt 2.17 **A1**

(c) Answer: Test statistic (with Yates' correction) = awrt 2.17

(d) df = 1 **B1**

(d) critical value = 3.841 **B1**

(d) valid conclusion ft: since 2.17 < 3.841, do not reject H0; insufficient evidence of an association between vaccination and catching flu **B1**

(d) Answer: df = 1, critical value = 3.841; 2.17 < 3.841 so do not reject H0.

© Revision Library | revisionlibrary.co.uk

Question 5

(a) H0: workshop attendance is independent of passing the resit exam **B1**

(a) H1: workshop attendance and passing the resit are associated (not independent) **B1**

(a) Answer: H0: workshop attendance and passing the resit are independent. H1: they are associated.

(b) uses E = (row total * column total) / grand total = 60*66/120 **M1**

(b) 33 cso **A1**

(b) remaining three values correct: 27, 33, 27 **A1**

(b) Answer: Expected frequencies: 33, 27, 33, 27.

(c) uses X^2 = sum of $(|O-E| - 0.5)^2 / E$ **M1**

(c) all four terms correct (unsimplified), e.g. 72.25/33, 72.25/27, 72.25/33, 72.25/27 **A1**

(c) cao, awrt 9.73 **A1**

(c) Answer: Test statistic (with Yates' correction) = awrt 9.73

(d) df = 1 **B1**

(d) compares 9.73 with the critical values 2.706 (10%), 3.841 (5%) and 6.635 (1%) **M1**

(d) 1%, since 9.73 > 6.635 **A1**

(d) Answer: df = 1; H0 is rejected even at the 1% level.

(e) strong evidence of an association between attending the workshop and passing the resit **B1**

(e) valid caveat, e.g. association does not imply causation; a confounding variable such as general motivation or prior attainment could affect both workshop attendance and passing, oe **B1**

(e) Answer: Strong evidence of association; however, a confounding variable (e.g. motivation) could explain the link without attendance causing the outcome.

© Revision Library | revisionlibrary.co.uk