2.6 Further Statistics: Hypothesis Testing and Chi-Squared Depth: 10-minute Pack

EDEXCEL 9FM0 · Calculator allowed · about 10 minutes

Total Marks

Name: _____ Date: ____ / ____ / _____

Answer ALL questions. Show all your working.

Key facts: chi-squared tests, in more depth

Original content written for Revision Library.

A chi-squared test compares observed frequencies O with the expected frequencies E predicted by a model, using X^2 = sum of (O-E)^2/E over all classes or cells, after any pooling. For a goodness-of-fit test to a fully specified distribution, df = (number of classes) - 1; if a parameter is estimated from the same sample (for example a Poisson mean, a binomial p, or a geometric p), one further degree of freedom is lost for each parameter estimated, because the estimation process artificially improves the apparent fit. For continuous data grouped into classes, expected frequencies are found from probabilities calculated using the standard Normal distribution, via z = (x - mu)/sigma and the function Phi. For an r by c contingency table testing independence between two categorical variables, E = (row total * column total)/grand total (this follows directly from the multiplication rule P(row and column) = P(row) * P(column) applied to the sample), and df = (r-1)(c-1). Classes or cells with an expected frequency below 5 should be pooled with a neighbouring class or column before the test statistic is calculated, and the degrees of freedom adjusted accordingly. Yates' continuity correction, X^2 = sum of (|O-E| - 0.5)^2/E, is applied only when df = 1 (every 2 by 2 contingency table), since this is the only case where the discrete test statistic is poorly approximated by the continuous chi-squared distribution. The signed standardised residual for a cell, e = (O-E)/sqrt(E), shows the direction as well as the size of a departure from the model, and the sum of its squares reproduces X^2. Because X^2 is never negative, only large (upper-tail) values ever provide evidence against H0; the calculated statistic is compared with the critical value from chi-squared tables at the chosen significance level and the stated degrees of freedom, and H0 is rejected if the statistic exceeds this value. As with any hypothesis test, the significance level equals P(Type I error), reducing it increases P(Type II error) for a fixed sample size, and increasing the sample size generally increases the power of the test; a chi-squared statistic also scales with sample size, so a very large sample can make even a small, unimportant deviation from a model 'statistically significant'. Combining ('pooling') data from distinct groups before testing can also mask or even reverse a genuine association found within each group, an effect known as Simpson's paradox.

© Revision Library | revisionlibrary.co.uk

1 Recall that two discrete random variables X and Y are said to be independent if P(X = x and Y = y) = P(X = x) * P(Y = y) for every possible pair of values x and y (the multiplication rule used, for example, when finding E(XY) for independent variables in FP.FS1).

(a) Explain how the null hypothesis of a chi-squared test for association between two categorical variables is connected to this definition of independence. **(2)**

(b) For a general 2 by 2 contingency table with row totals R1 and R2, column totals C1 and C2, and grand total n, the sample estimates of the marginal probabilities are P(row 1) = R1/n and P(column 1) = C1/n. Assuming independence, so that P(row 1 and column 1) = P(row 1) * P(column 1), show that the expected frequency for the (row 1, column 1) cell is exactly R1*C1/n, the usual formula E = (row total * column total) / grand total. **(2)**

(c) A student says: 'If the chi-squared test statistic works out to be exactly 0, the two variables must be perfectly independent.' Comment on whether this statement is accurate. **(2)**

(Total for Question 1 is 6 marks)

© Revision Library | revisionlibrary.co.uk

Mark scheme · 2.6 Further Statistics: Hypothesis Testing and Chi-Squared Depth: 10-minute Pack

Question 1

(a) H0 for a test of association states precisely that the two categorical variables are independent (oe) **B1**

(a) i.e. for every row category and column category, P(that row and that column) = P(that row) * P(that column) **B1**

(a) Answer: H0 is exactly the statement that the two variables are independent: P(row and column) = P(row) * P(column) for every cell.

(b) under independence, P(row 1 and column 1) = (R1/n) * (C1/n) **M1**

(b) cso: expected frequency = n * P(row 1 and column 1) = n * (R1/n) * (C1/n) = R1*C1/n, as required **A1**

(b) Answer: E(row 1, col 1) = n * (R1/n) * (C1/n) = R1*C1/n, matching the standard formula.

(c) a test statistic of 0 means every observed frequency in the sample exactly equals its expected (independence) value, so this sample gives the strongest possible evidence in favour of H0 **B1**

(c) however this only describes the particular sample collected; it does not prove that the underlying population variables are truly independent, since a different sample could show some discrepancy. It is evidence for, not proof of, independence **B1**

(c) Answer: Not fully accurate: a value of 0 gives the strongest possible sample evidence for independence, but it does not prove the population variables are independent.

© Revision Library | revisionlibrary.co.uk